கணிக்கோவை

கணித்தமிழ்24
மாநாட்டுக் கட்டுரைகள்

கணிக்கோவை

கணித்தமிழ்24

மாநாட்டுக் கட்டுரைகள்

சே.ரா. காந்தி, இ.ர.பா.ப.
இயக்குநர்

நாள்: 29.01.2024

முன்னுரை

தொழில்நுட்ப மாநாடுகளின் எழுத்துபூர்வ ஆவணமாக அம்மாநாட்டின் ஆய்வுக் கட்டுரைகள் திகழும். தகவல் தொழில்நுட்பத்துக்கு ஏற்பத் தன்னைத் தகவமைக்கும் தமிழ் மொழியின் வரலாற்றை, தொழில்நுட்பப் பயணத்தை, எதிர்கால சாத்தியத்தை விவாதிக்கும் பொருட்டு தமிழ்நாடு அரசால் பன்னாட்டுக் கணித்தமிழ்24 மாநாடு நடத்தப்படுகிறது. இம்மாநாட்டின் ஒரு பகுதியாக ஆய்வறிஞர்களிடமிருந்து ஆய்வுக் கட்டுரைகளைப் பெற்று அவற்றை விவாதிப்பது கணித்தமிழ் பயணத்தை மேம்படுத்துவதற்கு உதவியாக இருக்கும் என்ற கருத்து முன்வைக்கப்பட்டது. ஏனெனில், மொழியியல் அறிஞர்கள், தொழில்நுட்ப வல்லுநர்கள் உள்ளிட்டோர் உருவாக்கி அளிக்கும் அந்த ஆய்வுக் கட்டுரைகளில் புதிய கருத்துகளும் கண்டடைதல்களும் முதன்மை இடம்பிடிக்கும். அவை மொழியின் மேம்பாட்டுக்கும் புதிய முன்னெடுப்புகளுக்கும் தூண்டுகோல்களாக அமையும் என்பதால், ஆய்வறிஞர்களிடமிருந்து கட்டுரைகளைப் பெறும் முன்னெடுப்புகளை மேற்கொண்டோம். பெருந்திரள் மொழி மாதிரிகள், தமிழில் இயற்கை மொழி ஆய்வு, இயந்திர மொழிபெயர்ப்பு மற்றும் பன்மொழித் தொழில்நுட்பங்கள் உள்ளிட்ட பல்வேறு தலைப்புகளில் ஆய்வுக் கட்டுரைகளைப் பெற முடிவெடுக்கப்பட்டது. மொழியியல், தொழில்நுட்பம் உள்ளிட்ட துறையினர் அறியும் வகையில் இது தொடர்பான அறிவிப்புகள் விரிவாக விளம்பரப்படுத்தப்பட்டன.

ஆய்வறிஞர்களும் இந்த அறிவிப்பைக் கண்டு பேரார்வத்துடன் பங்களிக்க முன்வந்தனர். தமிழ் வளம் குன்றாததாகவும் தொழில்நுட்ப வளர்ச்சிக்கான சாத்தியம் கொண்டதாகவுமே தொடர்ந்து இருந்துவருகிறது என்னும் நம்பிக்கையை இந்த ஆர்வத்தில் உணர முடிந்தது. ஆய்வறிஞர் ஆய்வுக் கட்டுரைகளைச் சமர்ப்பிக்கும் முன்னர் அதன் சுருக்கத்தை அளிக்கக் கோரியிருந்தோம். அதன்படி, ஆய்வறிஞர்கள்

நூற்றுக்கும் மேற்பட்டோர் தாங்கள் எந்தப் பொருண்மையில் கட்டுரைகளைச் சமர்ப்பிக்க உள்ளோம் என்பதைச் சுருக்கமாக விளக்கும் வகையில் ஆய்வுக் கட்டுரைச் சுருக்கங்களை அனுப்பியிருந்தனர். அந்த ஆய்வுச் சுருக்கங்களை வல்லுநர் குழுவினர் முழுமையாக ஆய்ந்தறிந்து, கலந்தாலோசித்து எவரெவரிடமிருந்து ஆய்வுக் கட்டுரைகளைப் பெறலாம் என்பதை முடிவெடுத்தனர்.

வல்லுநர் குழு இறுதிப்படுத்திய ஆய்வறிஞர்கள் பல்வேறு பொருண்மைகளிலும் கட்டுரைகளை அளித்திருந்தார்கள். இந்தியாவின் புகழ்பெற்ற நிறுவனங்கள் இந்தக் கட்டுரைகளுக்குப் பங்களித்திருந்தன. அந்தக் கட்டுரைகளையும் வல்லுநர் குழு முழுமையாக ஆய்ந்தறிந்தது. அவற்றிலிருந்து தேர்ந்தெடுக்கப்பட்ட கட்டுரைகளைத் தொகுத்து கணிக்கோவை நூலாக்கியுள்ளோம். இந்தக் கணிக்கோவை நூலானது மொழியியல் ஆர்வலர்களுக்கும் ஆய்வு மாணவர்களுக்கும் ஆய்வறிஞர்களுக்கும் உதவியாக அமையும். இந்தக் கட்டுரைகள் விவாதிக்கும் பொருண்மைகள் தொடர்பான விவாதங்கள் ஆக்கபூர்வமான விளைவுகளை நோக்கி மொழியையும் மொழி ஆய்வையும் நகர்த்தும் என்பதில் ஐயமில்லை. ஆகவே, தமிழ்கூறும் நல்லுலகின் புது வரவான இந்த ஆய்வுக் கட்டுரைகளை மொழியியல் ஆர்வலர்கள் ஆழ்ந்து படித்து பயன் துய்க்க வேண்டும் என விரும்புகிறோம்.

நன்றி.

இயக்குநர்

CONTENTS

OPTICAL CHARACTER RECOGNITION FOR TAMIL

TAMIL CORPUS DEVELOPMENT AND LINGUISTIC RESOURCES

NATURAL LANGUAGE PROCESSING FOR TAMIL

SENTIMENT ANALYSIS AND EMOTION RECOGNITION IN TAMIL TEXT & SPEECH

TAMIL LANGUAGE TECHNOLOGIES FOR EDUCATION AND E-GOVERNANCE

TAMIL E-LEARNING PLATFORMS AND TOOLS

ROLE OF TAMIL IN SPATIAL COMPUTING

OPTICAL CHARACTER RECOGNITION FOR TAMIL

A Survey on Various Machine Learning Algorithms for the Translation of Tamil Scripts

Dr. E. Haripriya

ABSTRACT

The world has witnessed immense growth of the Tamil language from Thanitamil to Kanitamil since time immemorial, and the major contribution belongs to the ancestors and their records in the form of Kalvettu, Ollaisuvadi, Seppu Tagadu and numerous other ancient scriptures. Several Archaeological researches are done based on these ancestral records and evidences only. Since the Tamil language has evolved so many variations of its scripts from the 3rd century to present day, it is still hard to identify the script in the primeval records. Only few people are familiar with the techniques to read the old Tamil scripts such as Brahmi, Vatteluthu, Tamizhi and so on. It is essential to identify the scripts present in those ancient records and to know the values of ancestors and their life style, medicines, cultural values, arts and history. With the help of Artificial Intelligence and Machine learning, these can be made possible. This paper compares various machine learning algorithms and techniques used so far to translate the scripts present in the old records to modern Tamil language.

1. INTRODUCTION

Tamil Brahmi characters exist in between 300 BC and 100 AD. Ulaga Podhumarai Thirukural was written using Tamil Brahmi script. These Scripts serve as a gateway to know our ancient culture, civilization, Arts, Mathematics and Medicinal Values. Brahmi script is considered as the oldest ancient script in India. During the period of Ashoka, it becomes more popular. Mostly this script was found in the form of copper plates and rocks.

The Tamil Brahmi script was also called as Tamiri or Damiri and it was a varied from the South Indian Brahmi script. Old Tamil inscriptions were written using Tamil Brahmi. This script was used as the writing system in most of the regions in early days. They were found on caves, stones, poems, pots and coins. Both Tamil Brahmi inscriptions and Brahmi inscriptions are same in same places in the Indian subcontinent, such as the Ashoka Edict, but varied in few ways. Day by day these versions are varied and native vowels were followed. The Tamil Brahmi script was the parent script and from that later Vatteluthu was originated.

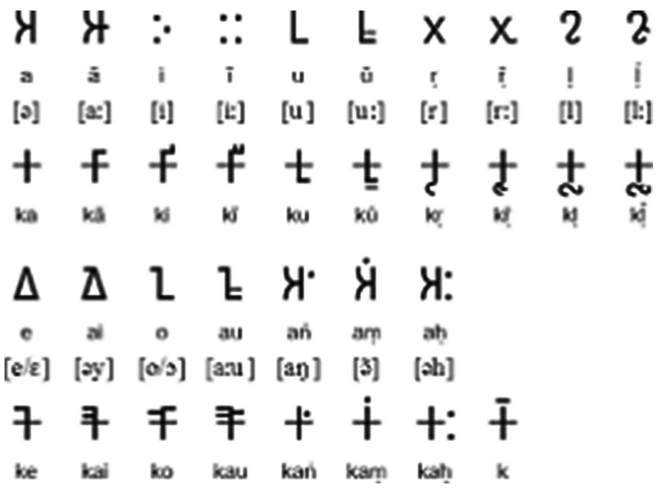


Fig 1: Brahmi Scripts - Alphabets



Fig 2: Brahmi Script in Copper Plates

Dr. E. Haripriya

Assistant Professor of Computer Science, J.K.K. Nataraja College of Arts & Science, Komarapalayam.

2. PROBLEM STATEMENT

Only few experts know these scripts and our Archaeological departments do these ancient script recognition with the help of these experts. Translation by Human experts is a time consuming process and may lead to errors.

3. LITERATURE REVIEW

Aniket et.al proposed an ensemble classification technique for the identification of touching Brahmi characters. The ensemble classifier is used to identify the touching and non-touching characters. Boosting algorithm is used for segmentation process. The proposed system has achieved the accuracy of 100% in identification and 99.16% for the segmentation of scripts.

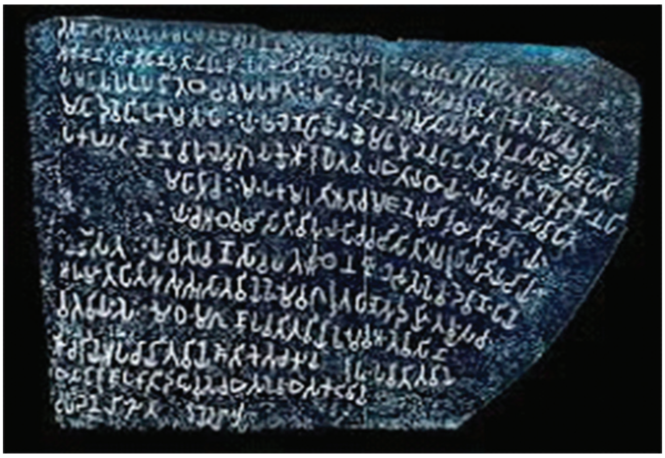


Fig 3: 3rd Century BCE Ashoka Rock edict, Chitradurga, Karnataka

Subadivya et.al suggested a model that uses Convolution Neural Network for extracting the features to translate the ancient Scripts and inscriptions into modern Tamil character. Here the dataset was built manually which consists the images. Flask framework was used and it achieves the accuracy of 94.6%.

Poornimathi et.al proposed a data augmentation technique for enhancing quality of the Inscriptions. Image Blur, Binarization and Edge detection techniques were used for the pre-processing the images. Edge detection was used to test the results.

Brindha et.al suggested a novel feature extraction technique for extracting the image features. NN Tool is used to train the featured images. The system achieved the accuracy of 91.6% and error rate of 8.7%.

Suganya et.al proposed a Fire safety algorithm for feature selection for the identification of ancient Tamil script. The system uses Shape and Hough transform for feature extraction and concatenation.

4. METHODOLOGY

To transfer the obtained image inscription into modern Tamil characters the following steps are involved. They are Image Pre-processing, Segmentation, Feature Extraction and Recognition. The pre-processing involves,

- **Image Inscription Input:** The images are captured using a high definition camera or scanner
- **Image Preprocessing:** For improving the quality of the input images, image preprocessing is done. It involves Data Collection, Cropping, Normalizing Image Inputs, Dimensionality Reduction and Data Augmentation
- **Segmentation:** To avoid processing of the entire image, the image is divided into segments and the needed segment is processed.
- **Feature Extraction:** The binarized image obtained is sliced to equivalent letter blocks each containing an ancient Tamil character.
- **Character Recognition:** The set of cropped images is fed to the Convolution Neural Network trained for Image Classification and recognition
- **Dataset:** To train the blocks of images from the available sample, a dataset is formed which is a combination of modern and ancient Tamil fonts.

Figure 4: Architecture of Character Recognition System

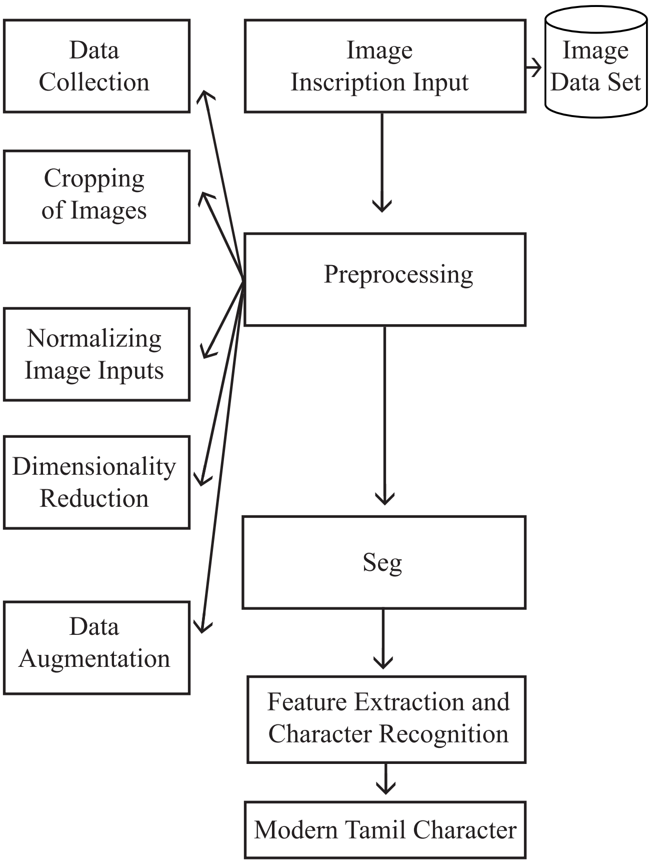


Table 1: Comparison table for various Machine Learning Methods

Author	Dataset gathering	Methodology Used	Library & Language Used	Efficiency Achieved
Aniket et.al	Cropped Images	Ensemble Classification and Boosting Algorithm	Python Library	Accuracy – 100% Segmentation – 99.16%
Brindha et.al	Images are captured from various places using high definition	Novel Feature Extraction & Zernike moment zoning feature cameras	Python Library	Accuracy – 92.14% Precision rate is poor
Poornimathi et.al	Cropped Images	Data Augmentation and Edge Detection	Python Library	Accuracy for Brahmi Scripts – 91.57% Vatteluthu – 89.75%
Subadivya.S et.al	Cropped Images	Natural Language Processing and CNN	Keras library with TensorFlow in BackEnd	Accuracy -94.6%
Suganya et.al	Cropped Images	Fire safety algorithm and Shape and Hough transform	Python	Accuracy -93.33%

CONCLUSION

This paper compares various Machine Learning models used for ancient Tamil script recognition. Most of the methods mainly concentrate on preprocessing of the images. By reading these studies, a significant rise has been achieved to extend the methods and standards. It is mainly used to compare various techniques used in the same field to improve the accuracy, precision and efficiency.

REFERENCES

1. Aniket S. Nagane & Sankar M. Mali, "Identification and Segmentation of Touching Brahmi Characters from Degraded Digital Estampage Images using Ensemble Classifier", Indian Journal of Computer Science and Engineering, Vol.12, No.6, 2021, pp-1722-1733.
2. Brindha S & Bhuvaneswari S, "Repossession and Recognition System: Transliteration of antique Tamil Brahmi typescript", Current Science, Vol.120, No.4, 2021, pp-654-665.
3. Poornimathi K, Muralibaskaran V and Priya L, "A Novel Pre-processing Technique for the Preservation of Tamil Brahmi Letters on Ancient Inscriptions in Different Application Domain", European Chemical Bulletin, Vol.12, No.10, 2023, pp-6372-6381
4. Sidhantha Poddar & Rohan Gupta, "Optical Script Identification for Multi Lingual Indic Script", Proceedings of IEEE International Conference, 2021
5. Subadivya S, Vigneswari J, Yamini M & Diviya M, "Script Character Recognition System using Deep Learning Technique", International Journal of Computer Science and Mobile Computing", Vol.9, No.6,2020, pp-114-119.
6. Suganya T.S, Murugavalli, "Feature selection for an automated ancient Tamil script classification system using machine learning techniques", IEEE International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies, 2017.

DHRITI - Multilingual OCR System: A Comparative Analysis of Tamil Document

Rajeev R R, Meharuniza Nazeem, Anitha R, Swathy A S, Dinesh Lal D L, Sabeerali K P

ABSTRACT

The research investigates the performance of three Optical Character Recognition (OCR) engines—Easy OCR, Paddle OCR, and Tesseract OCR—specifically focusing on their efficacy in handling Tamil printed text. The study aims to enhance the Document Handling and Retrieval of the Image Text Information (DHRITI) Multilingual system, by a comparative study of OCR engines available for Dravidian Languages. The recognition of Dravidian language documents is crucial as these invaluable records illuminate cultural, linguistic, and societal evolution, providing insights into ancient civilizations and preserving linguistic nuances within the Dravidian language family, enhancing their historical significance. Our focus here is on printed Tamil Document. After rigorous testing and analysis, Tesseract OCR emerged as the most effective solution for Tamil printed recognition. Tesseract recognizes new scripts with 96% accuracy and old scripts from the 1920s with 82% accuracy. The efficiency of Tesseract and its crucial function within the DHRITI architecture are highlighted in this research, which offers insightful information about optimizing OCR engines for Tamil scripts.

Rajeev R R, Meharuniza Nazeem, Anitha R, Swathy A S, Dinesh Lal D L, Sabeerali K P

International Centre for Free and Open Source Solutions (ICFOSS), Thiruvananthapuram.

*Corresponding author(s). E-mail(s): rajeev@icfoss.in, meharuniza@icfoss.org

INTRODUCTION

Optical Character Recognition is essential for transforming handwritten or printed text into machine-readable format in the age of the digital revolution. OCR plays a crucial role in preserving historical documents by converting them into digital formats. This technology facilitates efficient storage, retrieval, and reconstruction of historical records, ensuring their accessibility and preservation for future generations while enabling researchers to analyze and interpret these documents with ease. Low resource languages like Dravidian languages, preservation of historical data is crucial. In this work we are particularly focusing on Tamil documents.

Tamil scripts are known for their ancient origins, but they also have distinctive features that make it harder to distinguish them from printed documents. The Tamil language is distinguished by a variety of scripts, each with unique characteristics, such as the modern Tamil script and the classical Brahmi-based script.

Tamil scripts can be difficult to recognize since their characters are complex and contain both vowels and consonants. The existence of ligatures and compound characters, which call for specific recognition methods, adds even more complexity. Tamil scripts are characterised by their curvilinearity, which adds to their complex visual patterns. Accurate identification of these patterns requires the use of sophisticated Optical Character Recognition (OCR) techniques.

A layer of complication is added by Tamil's widespread usage of ligatures, which are combinations of numerous characters that create a single visual unit. These ligatures, like consonant-vowel pairings, require OCR algorithms that can precisely identify and interpret these complex structures.

Even with these obstacles, OCR is made more challenging by Tamil's contextual nature. Diverse typeface styles, dimensions, and handwriting philosophies add to the complexity of precisely extracting textual information from printed materials. Consequently, the efficiency of OCR techniques is critical for negotiating the intricacies of Tamil scripts and guaranteeing accurate recognition and extraction of textual data from a variety of sources.

This research concentrates on comparing three OCR engines to enhance the Document Handling and Retrieval of Image to Text Information (DHRITI) system, with Tesseract selected as the engine of choice for Tamil printed text recognition.

The following sections of this paper present a thorough analysis of modern OCR engines, emphasising their distinctive features. We explore which approaches are best suited to deal with the subtleties of the Tamil language.

We next take a closer look at the DHRITI-Multilingual OCR architecture and explain how Tesseract OCR is integrated as the recommended engine. The ensuing sections provide a thorough explanation of our analytical methodology, including the techniques used, the corresponding findings, and final observations.

RELATED WORKS

The presented works showcase diverse approaches to OCR specifically tailored for Tamil scripts, with a focus on ancient inscriptions and printed text recognition.

Lalitha Giridhar et al. [1] introduce a novel OCR method for old Tamil inscriptions in temples, spanning the 7th to 12th centuries. Employing image recognition-based categorization, they enhance OCR by incorporating Otsu thresholding for picture binarization. To generate audio output, a CNN connected to Tesseract with the Python Tesseract library and Google gTTS engine achieves an overall accuracy of 77.7% and a text-to-speech accuracy of 68%.

In 2002, K.G. Aparna and A.G. Ramakrishnan [2] propose an OCR system for Tamil language, specifically targeting printed text recognition. The classification process utilizes spatial occupancy, and the recognition process employs orthonormal transform features, achieving an impressive average recognition accuracy of 98%.

R. Jagadeesh Kannan et al. [3] conducted a comprehensive analysis of OCR for Tamil script in 2009. Their work covers Tamil script characteristics, the current state of OCR, and the methodologies employed. This paper serves as an overview of active research in OCR systems for Tamil scripts, providing a comparative examination.

M. Ramanan et al. [4] presented a hybrid decision tree method for SVM-based printed Tamil character identification in 2015. Utilizing a binary SVM arranged in a hybrid decision tree for multiclass classification, their approach, incorporating density, HOG, and transition properties, achieves a remarkable 98.8% identification rate.

In summary, these works collectively contribute innovative methods to address the unique challenges of OCR for Tamil scripts, spanning ancient inscriptions to contemporary printed text, and showcasing advancements in accuracy and methodology.

METHODOLOGY

OCR utilizes cutting-edge computer vision and Deep Learning techniques. OCR has advanced significantly and continues to be a pioneering challenge in deep learning and computer vision. Its many uses—from industrial applications to personal use to research and development—showcase its crucial importance in various fields.

The comparison investigation shows that Easy OCR, Paddle OCR, and Tesseract OCR have different strengths when it comes to identifying Tamil, English and Malayalam documents.

EasyOCR: It is based on PyTorch and Python and uses GPU acceleration to maximize performance for quick text recognition. The CRNN model, which consists of three primary components—ResNet for feature extraction, LSTM for sequence labeling, and CTC for decoding—manages recognition, while the CRAFT algorithm manages detection. This well-designed architecture improves the efficiency of the optical character recognition process. EasyOCR's minimum software dependencies make integration easier via its API and allow for simple usage without a lot of external requirements. This is one of its standout features. EasyOCR's [5] robust frameworks and algorithms, which put accessibility, speed, and accuracy first when identifying Tamil text in pictures, demonstrate the program's effectiveness.

Paddle OCR: PaddleOCR is an easy-to-use OCR toolbox that requires only a few lines of code to apply and train models [6]. It meets various needs with a range of pre-trained models, such as text detection and identification. Based on memory utilization, it provides models that strike a compromise between accuracy and speed. With support for more than 80 languages, our main product, PP-OCR, excels at English and Chinese recognition. With the use of CNNs, Bi-directional LSTMs, and a transcription layer, PaddleOCR's design ensures precise and effective OCR for printed Tamil texts. By using Connectionist Temporal Classification (CTC) Loss to improve training and decoding, a clear and accurate output is generated.

Tesseract: Several processes specifically designed to make use of the distinctive features of the script are required for Tesseract to recognize printed Tamil manuscripts. Using connected component analysis, the first preprocessing divides the image into separate parts and arranges them into lines and blobs. Taking into account the interconnectedness of Tamil characters, word segmentation is then used. With Tesseract [7],

recognition is done in two steps: the first pass tries basic recognition, and words that are recognized correctly are used as training for the second pass, which tries error correction. The spacing is fine-tuned, and small capital letters are identified [8]. Adding an LSTM model, the updated Tesseract analyses the input image line by line in rectangular boxes [9], improving sequential analysis-based Tamil script recognition. This modified approach guarantees precise segmentation and recognition for Tamil-printed documents, improving the OCR system's overall performance.

IMPLEMENTATION

The Tesseract engine, which is skilled at extracting text from papers, PDFs, or images, is smoothly integrated into our DHRITI - Multilingual OCR system. The architecture diagram explaining the setup and operation of the system is shown in the following section.

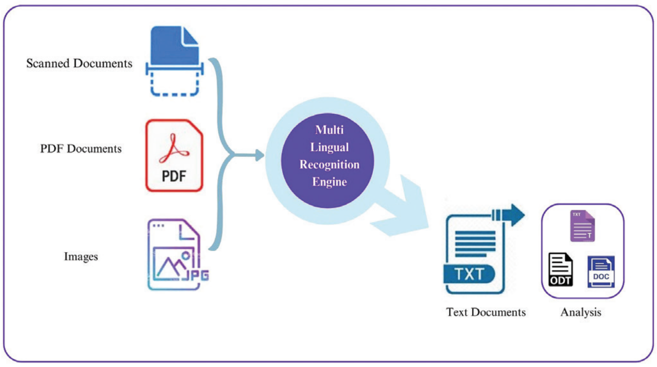


Fig.1: Architecture of the proposed system

We set out on a multilingual OCR to thoroughly investigate various approaches to document analysis, concentrating on materials from the 1920s as PDFs and images. This historical background broadens the scope of our analysis and illuminates a variety of linguistic subtleties. To evaluate the performance of OCR systems, we compare three major contenders across the language domains of Malayalam, Tamil, and English.

Our study's flexibility is in its ability to accept input in both image and PDF formats, which the OCR engine then carefully processes to extract text from the provided files. When the final output is displayed in text or document format, it prepares the reader for a thorough manual examination that examines the Character Error Rate (CER) to determine how accurate each result is. This painstaking analysis not only brings to light the intricacies of performance but also provides the foundation for a strong comparison that aims to identify the most ideal result.

Tesseract is an open-source program with multilingual compatibility that can help detect Tamil characters when configured with the appropriate language pack. The first step in the installation process is to integrate Tesseract OCR into the system before starting the extraction process. The pytesseract wrapper for Python makes it easier to communicate with the OCR engine. It is necessary to obtain and install the Tamil language pack in the Tesseract data directory. Preprocessing is required for the best OCR results when working with images. Converting the image to grayscale simplifies it, while contrast augmentation and binarization highlight text portions. Deskewing is used to fix alignment, and noise reduction is used to eliminate distortions. While cropping concentrates attention on the areas of interest, text localization finds pertinent sections. The image is further refined using optional color inversion and denoising. Normalization guarantees uniform values for pixels. The image to string function is used to deliver the processed image to Tesseract OCR, which is the central step in the process. Tamil text that has been extracted can then be saved, printed, or used as required.

There is an extra layer added for PDF files. Images are taken from each PDF page by traversing over them using the PyMuPDF library. Each image is thereafter subjected to the same preprocessing and OCR procedures, which together produce the extracted Tamil text from the whole PDF. Customization and adaptability are important factors. Optimizing language parameters, experimenting with preprocessing techniques, and fine-tuning configurations yield the best OCR results. Installing dependencies like pytesseract, Pillow, and fitz is necessary to enable the processing of images and PDFs. Tesseract OCR, Python libraries, and careful preparation techniques work together to provide a reliable method for removing Tamil text from pictures or PDF files.

EXPERIMENTAL ANALYSIS

This research paper explores the comparative performance of OCR engines as mentioned in previous sections within the DHRITI- Multilingual OCR system, emphasizing the enhancement of Tamil printed text recognition. Using a diverse dataset, the study assesses these engines based on Character Error Rate (CER) metrics, providing a nuanced evaluation of their accuracy.

The investigation reveals Tesseract OCR as the optimal choice for Tamil printed text recognition, highlighting its robust algorithms and language support. The paper delves into the integration of Tesseract OCR within the DHRITI system through a comprehensive block diagram, emphasizing its pivotal role in the document handling and retrieval process.

Results, depicted in a comparative performance graph, showcase Tesseract OCR's consistent outperformance. The research underscores the

importance of CER to capture the intricacies of Tamil text recognition. Ultimately, the findings contribute to informed decision-making in selecting OCR engines for effective integration into systems handling Tamil printed documents, thereby impacting document management frameworks like DHRITI.

We tested the models using the metrics CER. An OCR system's accuracy is measured by calculating the CER [5], which is determined by comparing the recognized text with a ground truth or reference text. The CER is calculated based on the number of character-level errors that the OCR system produces. The character mistake rate can be obtained using the following equation:

Whereas 'I' denotes character present in the recognised text but absent from the reference text, 'D' denotes deletions (characters missing from the recognised text compared to the reference text), 'S' denotes substitutions (characters recognised incorrectly), and 'N' denotes the total number of characters in the reference text.

In our analysis, the system undergoes testing with new input images or PDFs, and the extracted output is subsequently assessed against the original script for evaluation. To enable a comparison between two, the text must then be aligned at the character level for CER. The number of substitutions, deletions, and insertions is then used to calculate the errors. Substitutions represent incorrect elements in the OCR output; deletions represent elements that are present in the ground truth but absent from the OCR output; and insertions represent elements that the OCR system identified but are absent from the ground truth. Ultimately, the CER formulas are used to calculate the error rates, which yield a numerical representation of the accuracy of the OCR system. Reduced CER values signify better performance; these measures also help with iterative system modification by providing insights into particular error kinds.

A realistic Python implementation develops a set of functions to compute the CER, count the different sorts of errors, and align the ground truth and OCR output. While distinct functions count substitution, deletion, and insertion errors, the alignment function aligns words or characters in the ground truth and OCR output using dynamic programming or other alignment approaches. After that, the computed error rates are interpreted, providing insightful information about the system's pros and cons. This method offers a solid framework for assessing impartially how well OCR algorithms handle Tamil text extracted from pictures or PDFs.

The experiment analysis results are shown below. In the experiment, documents were sorted into two categories: old scripts, which were documents dated 1920 or earlier, and new scripts, which were documents labeled more recently.

OCR Engines	Accuracy		CER	
	Old Script	New Script	Old Script	New Script
Paddle OCR	0.6419	0.6637	0.35802	0.33628
Easy OCR	0.6823	0.7015	0.3177	0.2985
Tessearact	0.8238	0.9667	0.1761	0.03333

Table 1: Comparative analysis

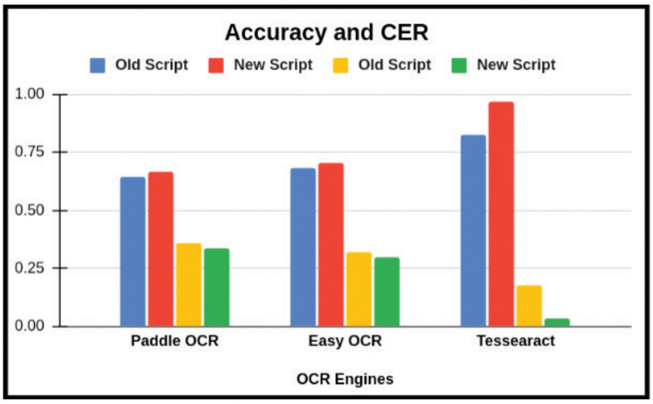


Fig.2: Comparative Performance Graph

Table 1 and Figure 2 demonstrate Tesseract's impressive performance in recognizing documents; in particular, it performs exceptionally well when understanding Tamil scripts in both its ancient and new versions. Our comprehensive investigation's results allow us to draw a certain conclusion: Tesseract performs better than other OCR techniques when it comes to Tamil script recognition, highlighting its superior efficacy and accuracy.

The DHRITI - Multilingual OCR is hosted and released by the ICFOSS infrastructure. A screenshot of the deployment can be found in Fig. 3. The outputs for the old and new scripts, respectively, are shown in figures Fig. 4 and Fig. 5.

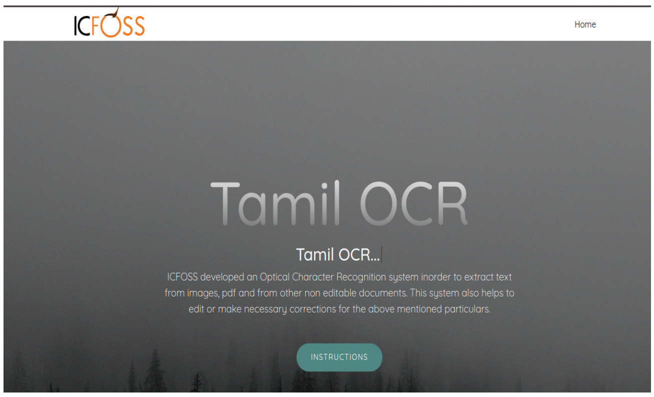


Fig.3: DHRITI - Multilingual OCR

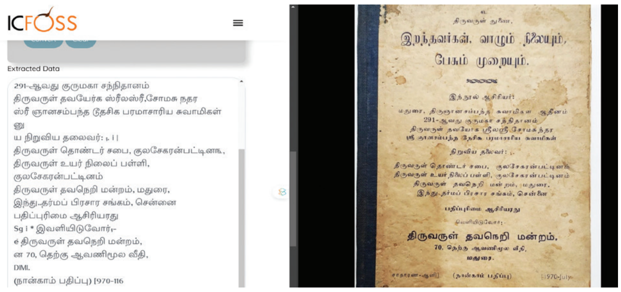


Fig.4: Output obtained for Old scripts

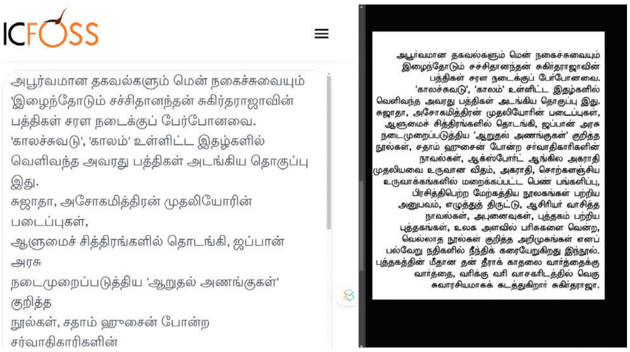


Fig.5: Output obtained for new scripts

CONCLUSION AND FUTURE WORK

In our work, we compare and assess the performance of three OCR engines: Tesseract, Paddle, and Easy OCR, in identifying printed Tamil documents using the DHRITI-Multilingual OCR system. Tesseract outperforms other transcription services when it comes to accurately transcribing printed Tamil text, as demonstrated by the analysis that includes metrics Character Error Rate (CER). Tesseract's architectural integration with DHRITI improves its functionality and helps to ensure that document processing and retrieval go smoothly.

The findings show that historical data performance has declined, with old script accuracy only reaching 82%. To improve performance, Tesseract must be retrained using old script documents as part of ongoing development. Furthermore, in order to increase accuracy, Tesseract will be retrained using a larger dataset that includes additional Dravidian languages, taking into account the unique linguistic peculiarities.

REFERENCES

- Giridhar, Lalitha, Aishwarya Dharani, and Velmathi Guruviah. "A novel approach to OCR using image recognition-based classification for ancient Tamil inscriptions in temples." arXiv preprint arXiv:1907.04917 (2019).
- Aparna, K. G., and A. G. Ramakrishnan. "A complete tamil optical character recognition system." Document Analysis Systems V: 5th International Workshop, DAS 2002 Princeton, NJ, USA, August 19–21, 2002 Proceedings 5. Springer Berlin Heidelberg, 2002.
- Kannan, R. Jagadeesh, and R. Prabhakar. "A comparative study of optical character recognition for Tamil script." European Journal of Scientific Research 35.4 (2009): 570-582.
- Ramanan, Muthulingam, Amirthalingam Ramanan, and Eugene Yougarajah Andrew Charles. "A hybrid decision tree for printed Tamil character recognition using SVMs." 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer). IEEE, 2015.
- D.R. Vedhaviyassh; R. Sudhan; G. Saranya; M. Safa; D. Arun, 2022 6th International Conference on Electronics, Communication and Aerospace Technology, "Comparative Analysis of EasyOCR and TesseractOCR for Automatic License Plate Recognition using Deep Learning Algorithm".
- Jaya Krishna Manipatruni1, R Gnana Sree, Ranjitha Padakanti, SreePriya Naroju4, Bharani Kumar Depuru Research Associate, Mentor, Team Leader, Research and Development, Director Innodatatics, Hyderabad, India, International Journal of Innovative Science and Research Technology ISSN No:-2456-2165, "Leveraging Artificial Intelligence for Simplified Invoice Automation: Paddle OCR-based Text Extraction from Invoices".
- R. Smith, Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), "An Overview of the Tesseract OCR Engine."
- A Complete Tamil Optical Character Recognition System K.G. Aparna and A.G. Ramakrishnan Biomedical Laboratory, Department of Electrical Engineering, Indian Institute of Science.
- Ujwala B S, Sumathi K, JNNCE Journal of Engineering & Management – A Peer Reviewed Bi-annual Journal Available online @ https://jjem.jnnce.ac.in ISSN: 2582-0079 (O), "A Novel Approach Towards Implementation Of Optical Character Recognition Using LSTM And Adaptive Classifier ."

Artificial Intelligence based Tamil Palm-Leaf Manuscript Reading Software

Balamurugan V T, Pravin Savaridass M, Udhaya Moorthy S J, Gokul S

ABSTRACT

Palm leaf manuscripts are an invaluable source of information containing literary, medical practices, religious, and historical texts in Tamil literature. However, modern scholars face difficulties in transcribing them due to the differences between the old script and modern Tamil characters, and the lack of experts who can transcribe them properly. To address this issue, an AI-powered software model is proposed to transcribe Tamil palm-leaf manuscripts into contemporary Tamil characters using image pre-processing, optical character recognition, and deep learning CNN model. The model is trained using a diverse dataset of palm leaf manuscripts, and convolutional neural networks are used to recognize the cursive Tamil characters in the manuscripts. The proposed approach achieves the model accuracy of 94% and 87% real time accuracy. The proposed AI-based Tamil palm leaf manuscript reader is much impactful in the world of cultural preservation and academic research, making it easier and faster to transcribe the huge collection of preserved and digitalized Tamil palm leaves.

1. INTRODUCTION

Palm leaf Manuscripts are really important because they help us understand the rich history and language of our societies. In Tamil literature, palm-leaf manuscripts are really important. Even though they're really valuable, modern scholars have a hard time transcribing them because of the differences between the old script and modern Tamil characters. Plus, there aren't many experts who can transcribe them properly, which makes it really hard to analyze and interpret them properly. Also, the average lifespan of a Palm leaf manuscript will be around 100 years after which they will start corroding.

Cultural heritage preservation is a high priority, and the Tamil language has been meticulously recorded on palm leaves for centuries due to its historical significance. These priceless informational treasures, appropriately called "Palm Leaf Manuscripts," are filled with a variety of historical, philosophical, and literary writings. But because of these manuscripts' delicate nature and the difficulties in digitizing and preserving them, creative solutions are now required.

We know how important it is to keep this history alive and make it available to more people, so we have proposed an AI-powered software model that leverages cutting-edge technologies to bridge the temporal gap between ancient Tamil script and its modern textual form. The task at hand involves transcribing Tamil palm leaf manuscripts using an AI-based model, integrating image pre-processing, optical character recognition (OCR), and deep learning, specifically utilizing Convolutional Neural Networks (CNNs). We've started by predicting the old Tamil numerals embedded in the palm leaf manuscript, and we've carefully curated a dataset to help us do this. As we go on, the goal is to expand the model's capabilities to cover the entire Tamil language, making it easier and faster to transcribe for the huge collection of preserved and digitalized Tamil palm leaves. Using AI and cutting-edge technology, this project is going to revolutionize the way scholars look at Tamil literature, making sure it's preserved and accessible for future generations. By quickly and accurately translating ancient scripts into a modern, easy-to-read format, this cutting-edge AI-based Tamil palm leaf manuscript reader is going to be a game-

changer in the world of cultural preservation and academic research.

2. LITERATURE SURVEY

This section outlines the contribution and other works done by various researchers for Palm leaf and other manuscript character identification in Tamil and other languages.

Ali and Joseph [1] developed a CNN model that is ideal for dispensing real-time input pictures that include Malayalam characters, as well as the task of segmenting words and typescripts from an image and predicting attractiveness using the CNN model. The gradient descent technique is used implicitly in CNN to perform feature extraction in this model. This technique is effectively used for digitizing Malayalam script, which consists of 36 consonants and 13 vowels, is approved out in stages, and has a training dataset accuracy of 97.26 percent.

By utilizing Simulated Annealing to optimize CNN, Balakrishnan and Pavithira [2] demonstrated the effectiveness of character appreciation. They talked about CNN's capabilities, various deep learning techniques, and CNN training methods. Character recognition, as defined, is the process of identifying and classifying characters in an input image and translating them into ASCII or another machine-readable format. The proposed method evaluated the OCR precision of multilingual texts from multiple books, and the results show that the CNN by SA has a higher accuracy than the unique CNN.

Alam Ahmad Hidayat et al [3] presented a study on Sundanese character recognition in Southeast Asian palm leaf manuscripts. The study focused on automating document image analysis for scanning ancient manuscripts using Convolutional Neural Networks (CNN). Two preprocessing strategies were investigated, with the second method, which used character binarization, beating the CNN-based classifier in the first technique, obtaining an astonishing 97.74% testing accuracy.

A Convolutional Neural Network (CNN) was proposed by Sabeenian et al. [4] with the impressive accuracy range of 96.1% to 100% for the recognition of Tamil palm-leaf characters. Effective feature extraction is the key to the CNN's performance, which is demonstrated by the 1000 samples per class for 15 classes in a dataset of scanned palm-leaf manuscript images.

Using machine learning and natural language processing, Cristea et al. [5] explore the evolution of optical character recognition (OCR) systems, from their early limits to their current capabilities. The shift towards Handwritten Text Recognition (HTR) and tools such as Monk and Transkribus are emphasized.

In a research paper written by Moudgil Aditi, et al [6] focus on handwritten Devanagari manuscripts, which are intricate because of a variety of features such characters, modifiers, and shirorekha. Deep learning is not fully exploited for Devanagari, despite its potential for character identification. The authors built a CapsNet character recognition system, preprocessed a Devanagari manuscript dataset, and addressed this issue. After that, they examined and contrasted the findings.

Singh et al [7] through their research proposes a new method for cleaning up old palm leaf manuscripts, which are important historical documents. The method combines two filters: a Gaussian filter to smooth out noise while keeping edges sharp, and a conservative smoothing filter to remove noise from the background without blurring the writing. The filters are applied to different parts of the image to get the best results. This method is shown to be effective in improving the quality of the manuscripts, making it easier to read and extract the text.

Based on [8] data augmentation approach, MAT-AGCA transforms the decoding of ancient Balinese lontar manuscripts. Combining variation creation and noise reduction, the technique achieves an amazing 96% accuracy in character identification, enabling automated interpretation and improving our grasp of Balinese cultural heritage.

The study by Islam et at [9] uses Image Data Generator to enhance a small dataset from the Electronic Beowulf manuscript. For various dataset augmentations, their proprietary Convolutional Neural Network (CNN) obtains recognition accuracies of 88.67%, 90.91%, and 98.86%. The CNN model also outperforms other algorithms with a benchmark accuracy of 99.03% on the MNIST dataset.

Balakrishnan Jayakumari et al. [10] conduct a comparative study on pretrained deep learning models for identifying Malayalam documents such as agreement contracts, notebook photos, and palm leaves. The refined VGG-16 model outperformed the competition, obtaining a high accuracy of 99.7%. Future work in creating algorithms for document classification based on content and spectral features is suggested in the paper.

3. MATERIAL AND METHODOLOGIES

3.1 Image Pre-processing

Preprocessing images is a crucial first step in the entire process, acting as a base for other steps. While working with various kinds of palm leaves that have different gradients, an optimized method is required. The process begins with the grayscale conversion of photos, an essential first step that lays the groundwork

for subsequent improvements. The subsequent conversion to binary is necessary but introduces a noise being exposed. In order to mitigate the effects of noise on the binary-converted images, strategic techniques are applied. The first technique is the use of contours and the fill-poly function. A crucial component of this technique is the use of adaptive thresholds, which are precisely calibrated to take into account variations in pixel values brought about by variations in lighting and image quality. In order to balance reducing noise and minimizing data loss, morphological operations such as the erode and dilate functions needs to be integrated. This extensive approach makes sure the best possible outcomes while displaying Adaptability in the face of evolving difficulties imposed by various palm leaves and external factors. As a result, image preprocessing emerges as a complex and essential step that sets the stage for subsequent steps of palm leaf manuscript the process of transcription.

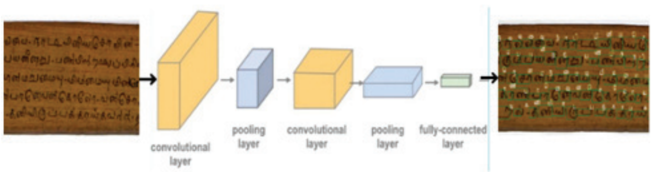
3.2 Optical Character Recognition (OCR)

The process of transcribing Tamil palm leaf manuscripts into contemporary Tamil script requires the use of Optical Character Recognition (OCR) technology. This method begins with digitizing fragile palm leaf manuscripts. Following that, OCR algorithms are used to detect individual characters from these images, splitting the written content into separate character for precise recognition. The predicted characters are then transcribed into contemporary Tamil script, boxed and labelled above the recognized characters to improve accessibility. This method not only preserves the cultural and historical significance of palm leaf manuscripts, but also makes them more accessible to researchers and the general public. It can be a useful tool for maintaining and sharing historical knowledge by implementing an evaluation step to guarantee accuracy before archiving the final transcribed text.

3.3 Deep Learning and Convolutional Neural Network (CNN)

The research uses a Convolutional Neural Network (CNN) without a separate feature extraction stage as shown in figure 1. Instead, all pixel values from a picture are immediately fed into the network's first layer. Tensor Flow is used for implementation, and a multilayer CNN architecture is used.

Figure 1: Block diagram of the proposed work



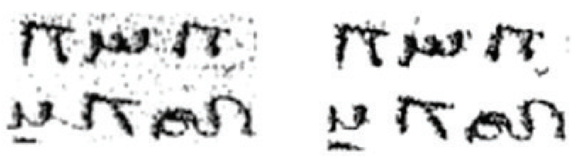
The first convolutional layer (C1) is crucial for processing visual data in the complex world of convolutional neural networks (CNNs). Consider a grayscale 50 x 50 pixel image being studied closely through a 3 x 3 pixel sliding window in this layer. The network can now extract important features and patterns from the input image through this process. The output then goes to a Rectified Linear Unit (ReLU) function after this convolutional operation. The ReLU functions play a crucial role in keeping training from turning saturating, which helps address the vanishing gradient issue and promotes improved retention. By strategically incorporating max-pooling procedures, the network's efficiency is further enhanced by a reduction in computational intensity. Depending on the network architecture, these pooling operations are spaced after each convolutional layer and are used to minimise the spatial dimensions, highlighting the most important features and improving computational speed at the same time. CNNs are built on this complex connection of convolution, activation, and pooling, which optimizes the ability of the system to recognise complexity and complex patterns in visual input.

4. RESULTS AND DISCUSSION

4.1 Noise Removal

Noise complicates the binarized image by introducing undesirable results. To overcome this issue, contour analysis, fillpoly functions, and morphological procedures like erode and dilate are routinely used to reduce noise while keeping vital information. A major part of image processing is balancing noise reduction without compromising detailed data. The below Figure 2 shows the obtained result.

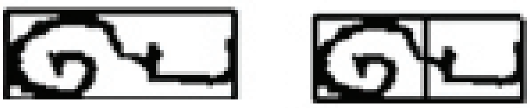
Figure 2: Noise Removal process



4.2 Segmentation

In the binary image, depending on the writer’s hand stroke, there have been a lot of joint characters. For the process of dataset creation and character prediction, these joint characters need to be separated. For this, we have used the maximum area separation method to create bounding boxes and separate the characters for data creation and model prediction. The figure 3 shows the obtained result.

Figure 3: Separated Joint characters



4.3 Classification and Recognition

The CNN model is used to recognize and classify the suggested palm leaf manuscript characters. The CNN algorithm instructs the computer to learn by doing. The CNN model was built and trained using the generated data set. The trained model has an accuracy of about 97%. When a palm leaf is given as an input, it will be pre-processed, segmented, each character will be classified and recognition will be done and it is depicted in figure 4, figure 5 and figure 6. The real time accuracy of the model is up to 87%.



Figure 4: Recognized characters in palm leaf

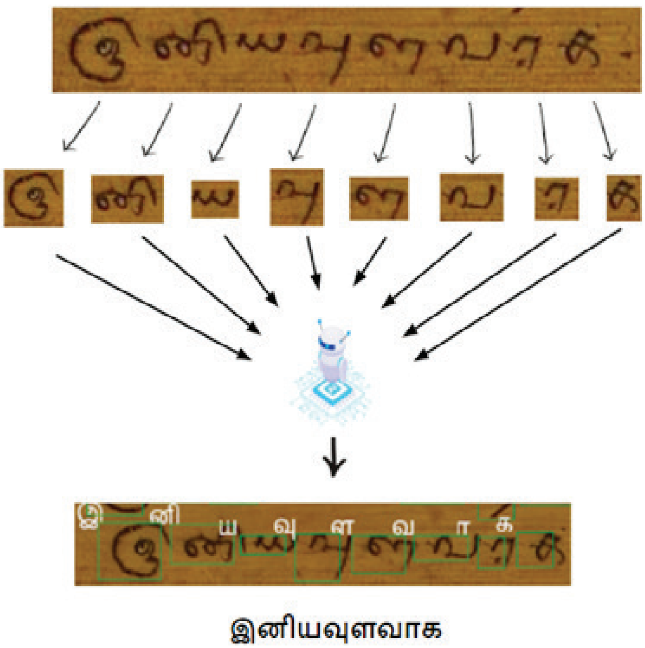


Figure 5: Recognition Process

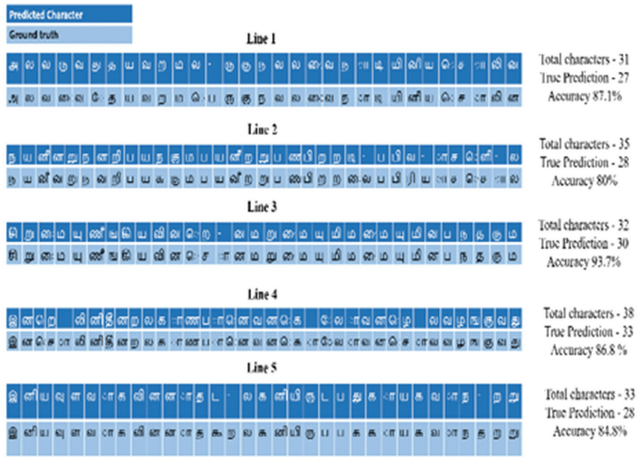


Figure 6: Accuracy analysis with ground truth

5. CONCLUSION

In this unique work, we present a cutting-edge AI-powered transcription tool designed especially for Tamil manuscripts. With the use of cutting-edge technologies like deep learning CNN, image processing, and optical character recognition (OCR), this tool can achieve a remarkable 94% model accuracy at an efficient 87% real-time accuracy. The convergence of these technologies facilitates transcription and acts as a vital connection between texts written in the ancient Palm leaf Manuscripts and contemporary eras. This study has major implications because it lays the foundation for a vital tool for academic research and cultural preservation. Even so, it's important to acknowledge that more testing and improvement are necessary in spite of the impressive results. With continued effort, the tool's dependability and adaptability should be strengthened, allowing for a smooth transition into the scholarly and cultural spheres. The wealth of knowledge that lies within Tamil heritage will ultimately be unlocked by this revolutionary technology, opening a dynamic door to a more thorough understanding of language evolution and cultural dynamics throughout history.

Limitations

Even though our suggested AI-based Tamil Palm leaf Manuscript Reader shows remarkable results, there are still some issues that need to be addressed, mainly with regard to image pre-processing. While our method works well for joint characters, it has difficulties with characters that are written wildly, especially those with long strokes, which could cause character separation errors. Furthermore, our pre-processing method's adaptiveness poses a challenge because different palm leaves require different image processing steps, even though it is necessary for different character formations. It introduces a practical limitation that each

bundle must have its processes modified accordingly. Our methodology is limited by the lack of a uniform model and inconsistent pre-processing techniques across various palm leaf collections. Acknowledging and resolving these limitations is essential to improving the tool's performance and guaranteeing its application for a wider variety of ancient Tamil manuscripts.

Ethics Statement

All the data required to work on this has been received from the Tamil Virtual Academy, various Tamil scholars, and activists. These digital copies of the palm leaf image have only been utilized for research purposes.

REFERENCES

[1] J. Ali and J. T. Joseph, "A convolution neural network-based approach for recognizing Malayalam handwritten characters," Malayalam Handwritten character recognition using cnn, vol. 9, no. 12, 2018.Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In Proceedings of the 24th International Conference on Machine Learning, pages 33–40.

[2] P. K. S. Balakrishnan and L. Pavithira, "Multi-font optical character recognition using Deep Learning," International Journal of Recent Technology and Engineering, vol. 8, 2019. James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation, 19(90):297–301.

[3] Alam Ahmad Hidayat, Kartika Purwandari, Tjeng Wawan Cenggoro, Bens Pardamean, "A Convolutional Neural Network-based Ancient Sundanese Character Classifier with Data Augmentation," Procedia Computer Science, Volume 179,2021Dan Gusfield. 1997. Algorithms on Strings, Treesand Sequences. Cambridge University Press, Cambridge, UK.

[4] Sabeenian, R.S., Paramasivam, M.E., Anand, R., Dinesh, P.M. (2019). "Palm-Leaf Manuscript Character Recognition and Classification Using Convolutional Neural Networks," Lecture Notes in Networks and Systems, vol 75. Springer, Singapore.

[5] Cristea, D., et al. "Bringing the Old Writings Closer to Us: Deep Learning and Symbolic Methods in Deciphering Old Cyrillic Romanian Documents." Memoirs of the Scientific Sections of the Romanian Academy 46 (2023).

[6] Moudgil, Aditi, et al. "Handwritten devanagari manuscript character's recognition using caps net." International Journal of Cognitive Computing in Engineering 4 (2023): 47-54.

[7] Singh, Mayank, and S. Indu. "Denoising of palm leaf manuscripts using Gaussian filter and conservative smoothing." AIP Conference Proceedings. Vol. 2521. No. 1. AIP Publishing, 2023.

[8] Sutramiani, Ni Putu, Nanik Suciati, and Daniel Siahaan. "MAT-AGCA: Multi Augmentation Technique on small dataset for Balinese character recognition using Convolutional Neural Network." ICT Express 7.4 (2021): 521-529.

[9] Islam, Mohammad Anwarul, and Ionut E. Iacob. "Manuscripts Character Recognition Using Machine Learning and Deep Learning." Modeling 4.2 (2023): 168-188.

[10] Balakrishnan Jayakumari, Bipin Nair, and Amel Thomas Kavana. "Classification of heterogeneous Malayalam documents based on structural features using deep learning models." International Journal of Electrical & Computer Engineering (2088-8708) 13.1 (2023).

TAMIL CORPUS DEVELOPMENT AND LINGUISTIC RESOURCES

செவ்வியல் தமிழ் சொற்பிரிப்புக் கருவி

முனைவர் இரா.அகிலன்

ஆய்வுச்சுருக்கம்

இயற்கை மொழி ஆய்வு என்பது இயற்கை மொழி அமைப்பைக் கணினிக்கு ஏற்ற வகையில் இலக்கணமாகக் கொடுப்பது. கணினிக்கு இயற்கை மொழி ஆய்வை மேற்கொள்ளும் அறிவுத்திறனைக் கொடுப்பதற்குப் பல மொழிப் பயன்பாட்டுப் பணிகளை மேற்கொள்ள வேண்டும். இதில் மிக அடிப்படையானது தரவக உருவாக்கம். கணினி புரிந்துகொள்ளும் வகையிலான தரவகம் உருவாக்கப்பட வேண்டும். மேலும் உருவாக்கப்பட்ட தரவகத்தை அடிப்படையாகக் கொண்டு மொழிப் பயன்பாட்டு ஆய்வுகளுக்கான மொழிக் கருவிகள் உருவாக்கப்பட வேண்டும். செவ்வியல் தமிழ் சொற்பிரிப்புக் கருவி என்பது உள்ளீடு செய்யப்பெற்ற செவ்வியல் பாடல்களை ஏற்கெனவே உருவாக்கப்பட்ட தரவுகளை அடிப்படையாகக் கொண்டு மூலப்பாடத்திலிருந்து சந்தி பிரித்த பாடம், சொற்கள் பிரித்த பாடம் எனச் சங்க இலக்கியங்களைக் கணினி மொழியியல் ஆய்வுகளுக்குத் தேவையான தரவுகளாக உருவாக்கி அளிக்கக்கூடிய மென்பொருள். செவ்வியல் தமிழ் சொற்பிரிப்புக் கருவி உருவாக்கம், தரவக உருவாக்கத்தில் இக்கருவியின் பங்கு, சொற்பிரிப்புக் கருவி பயன்பாடுகள், கருவி உருவாக்கத்தில் ஏற்படும் சிக்கல்கள் மற்றும் தீர்வுகள் பற்றி விரிவாக ஆராய்வதே இதன் நோக்கமாக அமைகிறது.

முனைவர் இரா.அகிலன்
நிரலாளர், செம்மொழித் தமிழாய்வு மத்திய நிறுவனம், சென்னை.

1. அறிமுகம்

ஒரு குறிப்பிட்ட ஒழுங்குமுறையுடன் அதிக அளவில் தேர்வு செய்யப்பட்டுச் சேமிக்கப்பட்ட இயற்கையான நடைகளை உடைய உரைகள் மற்றும் பல்வேறு பனுவல்களின் தொகுப்பு பெருந்தரவு எனப்படும். இஃது ஒரு குறிப்பிட்ட மொழியின் சொற்கள் அல்லது துறைச் சொற்களை உள்ளடக்கியதாக இருக்க வேண்டும். ஒரு மொழியின் பல்வேறு பரிணாமங்களை எதிரொலிப்பதாகவும் பலதரப்பட்ட புலங்களுக்கு முதன்மை அளிப்பதாகவும் அறிவியல் முறைப்படியும் இருக்க வேண்டும். தரவகம் என்பது பல்வேறு வகைப்படும். ஒரு மொழியின் வரலாற்று வளர்ச்சியைக் காண்பதற்கான தரவகம் வரலாற்றுமுறைத் தரவகம் (Historical Corpus) என்றும், ஒரு குறிப்பிட்ட மொழியின் இன்றைய அமைப்பை அல்லது இலக்கணத்தைப் பெறுவதற்கான தரவகம் ஒத்திசைவுத் தரவகம் (Synchronic Corpus) என்றும், மொழி கற்பித்தல், கற்றலுக்கான தரவகம் (Corpus for Language Learning/Teaching), என்றும் இயந்திர மொழிபெயர்ப்புகளுக்குப் பயன்படும் இரண்டு மொழிகள் அல்லது அதற்கு மேற்பட்ட மொழிகளின் இணைகளை உள்ளடக்கிய இணைத் தரவகம் (Parallel Corpus) என்றும் பல வகைகளில் பகுக்கப்படுகின்றன. இவ்வாறு பல்வேறு புலங்களுக்குப் பயன்படும் வகையில் தரவகங்களை உருவாக்கப் பல்வேறு வகையான மொழியியல் ஆய்வுக் கருவிகள் பயன்படுகின்றன. இதில் செவ்வியல் தமிழ் சொற்பிரிப்புக் கருவி மூலப்பாடத்திலிருந்து சந்தி பிரித்த பாடம் மற்றும் சந்தி பிரித்த பாடத்திலிருந்து சொற்கள் பிரித்த பாடம் எனச் சங்க இலக்கியங்களைக் கணினி மொழியியல் ஆய்வுகளுக்குத் தேவையான தரவுகளாக உருவாக்கி அளிக்கிறது. இதில் பெறப்படும் தரவுகள் இயற்கை மொழி ஆய்வில் செவ்வியல் நூல்களின் பல்வேறு பரிமாணங்களையும் ஆய்வதற்கு அடிப்படை ஆதாரமாக அமைகிறது.

2. செவ்வியல் தமிழ்த் தரவகம்

தமிழ் மொழிக்கான தரவகம் சங்க இலக்கியம், பக்தி இலக்கியம், தற்காலத் தமிழ் என மூன்று வகையான தரவகமாக உருவாக்கப்பட வேண்டும். தமிழில் உள்ள நாற்பத்தொரு செவ்வியல் நூல்கள் கி.மு 3 ஆம் நூற்றாண்டு முதல் கி.பி 6 ஆம் நூற்றாண்டு வரையிலான காலகட்டத்தைச் சேர்ந்தது. பின்னர் கி.பி 7

ஆம் நூற்றாண்டு முதல் கி.பி 9 ஆம் நூற்றாண்டு வரை இயற்றப்பட்ட இலக்கியங்கள் பக்தி இலக்கிய வகையைச் சார்ந்தது. கி.பி 10 ஆம் நூற்றாண்டு முதல் தற்போது வரையுள்ள மொழி தற்காலத் தமிழாகப் பயன்பாட்டில் இருந்து வருகிறது.

இந்த மூன்று வகையான காலகட்டத்தில் தமிழ் மொழியின் சொற்கள், பயன்பாடு பல்வேறு வகைகளில் மாறுபட்டு உள்ளது. எனவே தமிழ் மொழிக்கான தரவகம் உருவாக்கும்போது இவற்றைக் கருத்தில் கொண்டு உருவாக்கப்பட வேண்டும். செவ்வியல் தமிழ் நூல்கள் 41 (தொல்காப்பியம், பத்துப்பாட்டு, எட்டுத்தொகை, பதினெண் கீழ்க்கணக்கு, சிலப்பதிகாரம், மணிமேகலை, இறையனார் களவியல், முத்தொள்ளாயிரம்) நூற்பாக்கள் வடிவிலும், பாடல் வடிவிலும் இடம்பெற்றுள்ளன. செவ்வியல் தமிழ் நூல்களுக்கு எனப் பல்வேறு தரவக உருவாக்க முயற்சிகள் நடைபெற்று வந்தாலும் சங்க இலக்கியங்களை மூலப்பாடம், சந்தி பிரித்த பாடம், சொற்கள் பிரித்த பாடம் என வகைப்படுத்திக் கணினி புரிந்துகொள்ளும் வகையிலும் இயற்கை மொழி ஆய்விற்கு உட்படுத்தும் வகையிலும் தரவகம் உருவாக்கப்பட வேண்டும். சென்னையில் உள்ள செம்மொழித் தமிழாய்வு மத்திய நிறுவனம் செவ்வியல் நூல்களுக்கான தரவகங்களை இந்த மூன்று நிலையில் உருவாக்கி உள்ளது.



இடம்பெற்ற அடிகள் : 18			மொத்த அடிகள் : 37543		
மூலபாடம்	சந்திபிரித்தபாடம்	சொற்கள்பிரித்தபாடம்	நூல்	பாடல்எண்	அடி
மிகு கடறருமணியொடு முதநியாததநோணி நெறிசெறி	மிகு கடறரு மணியொடு முத்து யாத்தநோணி நெறிசெறி	மிகு கடல் அரு மணி ஒடு முத்து யாத்த நோணி நெறி செறி	பரி	1	16
நின்னஞ்சிக் கடற்பாய்ந்த பிணிநெகிழ் பவிழ்தண்டார்	நின் அஞ்சிக் கடல் பாய்ந்த பிணி நெகிழ்பு அவிழ் தண் தார்	நின் அஞ்சி கடல் பாய்ந்த பிணி நெகிழ்பு அவிழ் தண் தார்	பரி	3	55
பாயிரும் பனிக்கடல் பார்துகள் படப்புக்குச்	பாய் இரும் பனிக் கடல் பார் துகள் படப் புக்குச்	பாய் இரும் பனி கடல் பார் துகள் பட புக்கு	பரி	5	1
நிறைகடன் முகந்துராய் நிறைந்துநீர் துளும்புந்தம்	நிறை கடல் முகந்து உராய் நிறைந்து நீர் துளும்பும் தம்	நிறை கடல் முகந்து உராய் நிறைந்து நீர் துளும்பும் தம்	பரி	6	1
காலொடு மயங்கிய கலிழ்கடலென	காலொடு மயங்கிய கலிழ் கடல் என	கால் ஒடு மயங்கிய கலிழ் கடல் என	பரி	8	31
மால்கடல் குடிக்கு மழைக்குரலென	மால் கடல் குடிக்கும் மழைக் குரல் என	மால் கடல் குடிக்கும் மழை குரல் என	பரி	8	32

படம் 1: செவ்வியல் தமிழ்த் தரவகம்

இந்தத் தரவு செவ்வியல் ஆய்வின் அனைத்துக் கூறுகளையும் இயற்கை மொழி ஆய்வில் உட்படுத்தக்கூடிய அடிப்படைத் தரவாக அமையும்.

3. இயற்கை மொழி ஆய்வுக் கருவிகள்

இயற்கை மொழி ஆய்வில் இந்திய மொழிகளுக்கான பயன்பாடு தன்னிறைவை அடைய முழுமையான தரவகம், குறியீட்டுத் தரவகம் மற்றும் மொழியியல் ஆய்வுக் கருவிகள் உருவாக்கப்பட வேண்டும். இவை உரைத் தரவகம், குறியீட்டுத் தரவகம், பேச்சுத் தரவகம், உரை சுருக்கி, சொல் வங்கி, தேடுபொறி, இயந்திர மொழிபெயர்ப்பு, மின் அகராதிகள், உருபனியல் பகுப்பான், சந்தி திருத்தி, சொல் திருத்தி, இலக்கணத் திருத்தி போன்ற கருவிகள் அடிப்படையானவை. மத்திய அரசின் தகவல் தொழில்நுட்ப அமைச்சகம் இதற்கென்ற ஒரு தனித் திட்டத்தை, இந்திய மொழிகளில் தொழில்நுட்ப வளர்ச்சி (Technology Development of Indian Language - TDIL) என்ற திட்டத்தை உருவாக்கிச் செயல்பட்டு வருகிறது. தமிழகத்திலும் மாநில அரசானது தமிழ் மொழியின் தொழில்நுட்ப வளர்ச்சிக்காகப் பல்வேறு திட்டங்களைச் செயல்படுத்தி வருகிறது. மேலும், தமிழ் இணையக் கல்விக்கழகம், சென்னை, ஹைதராபாத் பல்கலைக்கழகம் தெலங்கானா, அண்ணா பல்கலைக்கழகம் கணிப்பொறியியல் துறை, சென்னை குரோம்பேட்டை AU-KBC மையம், அண்ணாமலைப் பல்கலைக்கழகம் சிதம்பரம், தஞ்சை தமிழ்ப் பல்கலைக்கழகம், கோவை பாரதியார் பல்கலைக்கழகம், இந்தியத் தொழில்நுட்பக் கழகம், அமிர்தா பல்கலைக்கழகம் கோயம்புத்தூர், இந்திய அறிவியல் கழகம், பெங்களூரு, மைக்ரோ சாப்ட்வேர், பெங்களூரு முதலான நிறுவனங்கள் தமிழ் மொழிக்கான தொழில்நுட்பக் கருவிகள் உருவாக்கத்தில் தொடர்ந்து ஈடுபட்டு வருகின்றன. மத்திய அரசின் கல்வி அமைச்சகத்தின் கீழ் இயங்கிவரும் செம்மொழித் தமிழாய்வு மத்திய நிறுவனம் நாற்பத்தொரு செவ்வியல் நூல்களுக்கான தரவகம் மற்றும் செவ்வியல் நூல்களுக்கான மொழி ஆய்வுக்கருவிகள் உருவாக்கத்தில் ஈடுபட்டு வருகிறது.

4. செவ்வியல் சொற்பிரிப்புக் கருவி

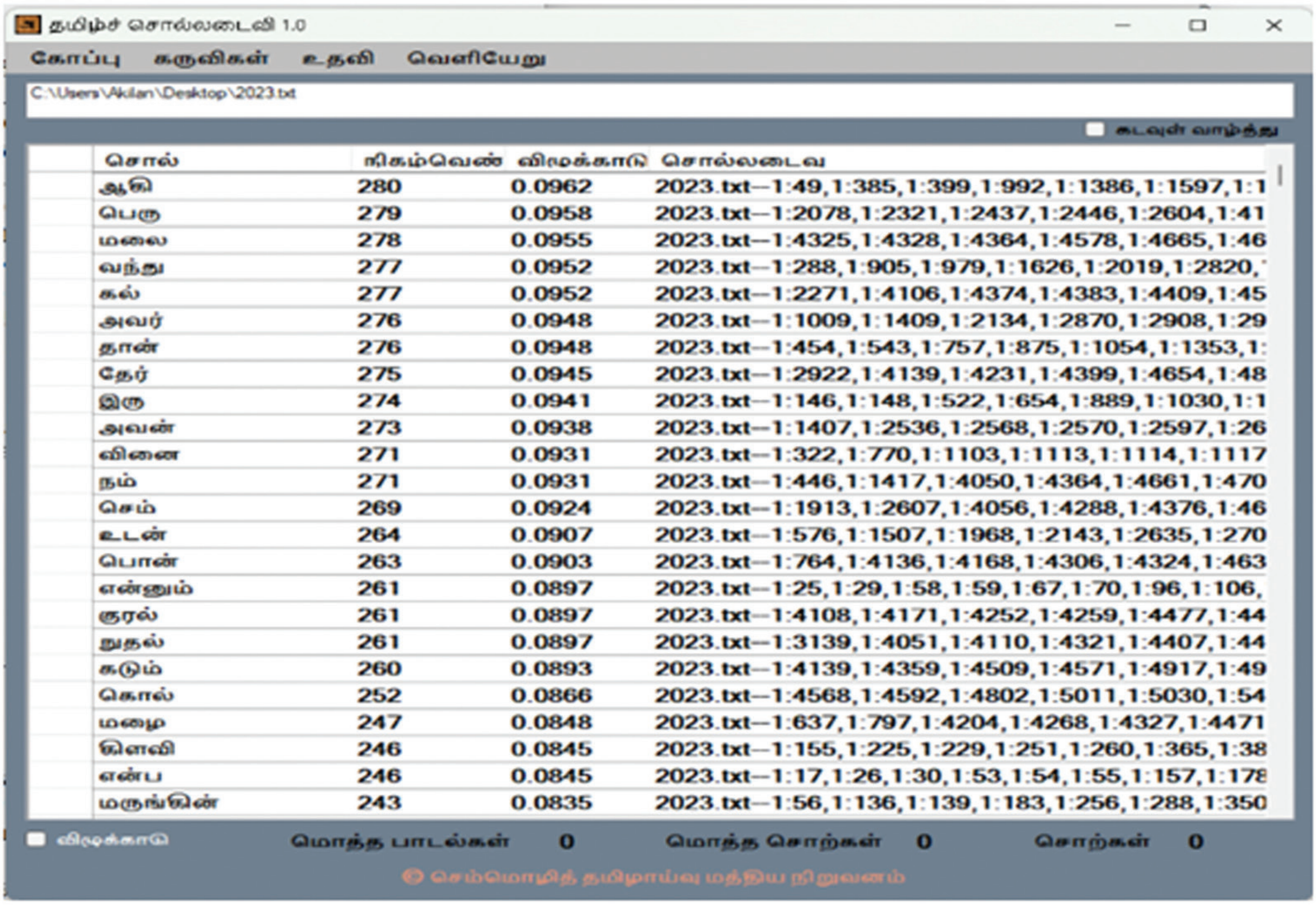
தமிழ் மொழி ஆய்வுக்கு உருவாக்கப்படும் மொழி ஆய்வுக் கருவிகள் சங்க இலக்கியத் தரவுகளை முழுமையாகப் பயன்படுத்தி ஆய்வுகள் மேற்கொள்ள முடியாது. செவ்வியல் தமிழ் நூல்கள் நூற்பாக்கள் வடிவிலும், பாடல் வடிவிலும் அமைந்துள்ளதால், தற்காலத் தமிழுக்கென உருவாக்கப்படும் கருவிகளைச் செவ்வியல் தரவக ஆய்வுகளுக்கு உட்படுத்துவதில் சிக்கல் ஏற்பட்டுள்ளது. ஆனால் செவ்வியல் தரவுகளை வேர்ச்சொற்கள் இனம் காணல், சங்க இலக்கியத் தரவுதளம், உரை ஆய்வு, தேடுபொறி, அகராதி, குறியீட்டுத் தரவகம் போன்றவற்றை மேற்கொள்வதற்கு இவற்றை மூலப்பாட வடிவில் உள்ள பாடல்களையும், நூற்பாக்களையும் உரை வடிவிலும், தனிச் சொற்களாகவும் பிரிக்க வேண்டிய தேவை உள்ளது. செவ்வியல் தமிழ் சொற்பிரிப்புக் கருவி என்பது கொடுக்கப்பட்ட உள்ளீடு செய்யப்பெற்ற சங்கப்பாடல்களை ஏற்கெனவே உருவாக்கப்பட்ட தரவுகளை அடிப்படையாகக் கொண்டு மூலப்பாடத்திலிருந்து சந்தி பிரித்த பாடம், சொற்கள் பிரித்த பாடம் எனச் சங்க இலக்கியங்களைக் கணினி மொழியியல் ஆய்வுகளுக்குத் தேவையான தரவுகளாக உருவாக்கி அளிக்கக்கூடிய மென்பொருள். ஏற்கெனவே உருவாக்கப்பட்ட தரவகம், செவ்வியல் தமிழ் வேர்ச்சொற்கள் மற்றும் கணினிக்கான விதிகள்

இவற்றை அடிப்படையாகக் கொண்டு இக்கருவி உருவாக்கப்பட்டுள்ளன.

5. சொற்பிரிப்புத் தரவகம் உருவாக்கம்

5.1 தரவு உருவாக்கம்

சொல் பிரிப்புக் கருவியின் நிரலாக்கத்தின் அடிப்படைத் தேவை தரவுகள் உருவாக்கம். இக்கருவி விரைவாகச் செயல்பட்டு முடிவுகளைத் தருவதற்கு உயர் அதிர்வெண் வரிசையில் தரவுகள் உருவாக்கப்பட வேண்டும். ஏற்கெனவே உருவாக்கப்பட்ட தரவுகளை நிகழ்வெண்களை அடிப்படையாகக் கொண்டு அவற்றின் முடிவுகளை உயர் அதிர்வெண் வரிசையில் கணினி புரிந்து கொள்ளும் வகையில் நிரலாக்கம் செய்ய வேண்டும். இதற்குத் தரவுகளை உள்ளீடு செய்தால் அவற்றைப் பிரித்து உயர் அதிர்வெண் வரிசையில் சொற்களைப் பிரித்து அளிக்கக்கூடிய மென்பொருளைப் பயன்படுத்தி அதற்கான பட்டியலைப் பெறலாம். செம்மொழித் தமிழாய்வு மத்திய நிறுவனம் உருவாக்கி வெளியிட்டுள்ள தமிழ்ச் சொல்லடைவு மென்பொருள் இதற்குப் பயன்படுத்தப்பட்டு நிகழ்வெண் வரிசையிலான தரவுகள் உருவாக்கப்பட்டது.



சொல்	நிகழ்வெண்	விழுக்காடு	சொல்லடைவ
ஆகி	280	0.0962	2023.txt—1:49,1:385,1:399,1:992,1:1386,1:1597,1:1
பெரு	279	0.0958	2023.txt—1:2078,1:2321,1:2437,1:2446,1:2604,1:41
மலை	278	0.0955	2023.txt—1:4325,1:4328,1:4364,1:4578,1:4665,1:46
வந்து	277	0.0952	2023.txt—1:288,1:905,1:979,1:1626,1:2019,1:2820,
கல்	277	0.0952	2023.txt—1:2271,1:4106,1:4374,1:4383,1:4409,1:45
அவர்	276	0.0948	2023.txt—1:1009,1:1409,1:2134,1:2870,1:2908,1:29
தான்	276	0.0948	2023.txt—1:454,1:543,1:757,1:875,1:1054,1:1353,1:
தேர்	275	0.0945	2023.txt—1:2922,1:4139,1:4231,1:4399,1:4654,1:48
இரு	274	0.0941	2023.txt—1:146,1:148,1:522,1:654,1:889,1:1030,1:1
அவன்	273	0.0938	2023.txt—1:1407,1:2536,1:2568,1:2570,1:2597,1:26
வினை	271	0.0931	2023.txt—1:322,1:770,1:1103,1:1113,1:1114,1:1117
நம்	271	0.0931	2023.txt—1:446,1:1417,1:4050,1:4364,1:4661,1:470
செம்	269	0.0924	2023.txt—1:1913,1:2607,1:4056,1:4288,1:4376,1:46
உடன்	264	0.0907	2023.txt—1:576,1:1507,1:1968,1:2143,1:2635,1:270
பொன்	263	0.0903	2023.txt—1:764,1:4136,1:4168,1:4306,1:4324,1:463
என்னும்	261	0.0897	2023.txt—1:25,1:29,1:58,1:59,1:67,1:70,1:96,1:106,
குரல்	261	0.0897	2023.txt—1:4108,1:4171,1:4252,1:4259,1:4477,1:44
நுதல்	261	0.0897	2023.txt—1:3139,1:4051,1:4110,1:4321,1:4407,1:44
கடும்	260	0.0893	2023.txt—1:4139,1:4359,1:4509,1:4571,1:4917,1:49
கொல்	252	0.0866	2023.txt—1:4568,1:4592,1:4802,1:5011,1:5030,1:54
மழை	247	0.0848	2023.txt—1:637,1:797,1:4204,1:4268,1:4327,1:4471
ஔவி	246	0.0845	2023.txt—1:155,1:225,1:229,1:251,1:260,1:365,1:38
என்ப	246	0.0845	2023.txt—1:17,1:26,1:30,1:53,1:54,1:55,1:157,1:178
மருங்கின்	243	0.0835	2023.txt—1:56,1:136,1:139,1:183,1:256,1:288,1:350

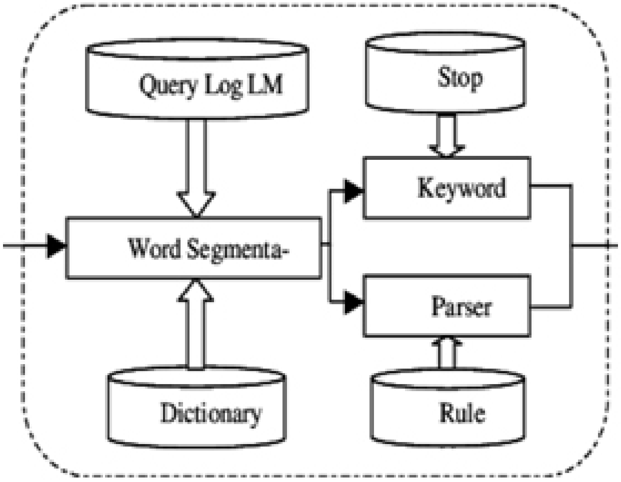
படம் 2: உயர் அதிர்வெண் தரவு உருவாக்கக் கருவி

மேற்கண்ட மென்பொருளில் ஏற்கெனவே சந்தி பிரித்த பாடத்தை உள்ளீடு செய்து அதற்கான பட்டியலைக் கருவிகள் என்னும் பகுதியில் சொல்லடைவு என்பதைத் தெரிவு செய்யும்போது அதற்கான தரவுகள் கிடைக்கும். இதை ஒரு தரவாக உருவாக்கிக்கொள்ள வேண்டும், பின்னர் மூலப்பாடத் தரவுகளை இக்கருவியில் உள்ளீடு செய்து அதற்கான உயர் அதிர்வெண் வரிசையில் சொற்களைத் தயார் செய்ய வேண்டும். பின்னர் இந்த இரண்டு தரவுகளையும் ஒப்பீடு செய்து அவற்றை இணைத்தரவுப் பட்டியலாக மாற்றி அமைக்க வேண்டும். இந்தத் தரவு இக்கருவி உருவாக்கத்திற்கு அடிப்படை ஆதாரமாக அமைகிறது.

5.2 சொற்பிரிப்புக் கருவி உருவாக்கம்

சொற்பிரிப்புக் கருவியின் உருவாக்கத்திற்குத் தரவு உயர் நிகழ்வெண் அடிப்படையில் வடிவமைக்கப்படுகிறது. கணினி சங்க இலக்கியச் சொற்களைப் பிரிக்கும்போது

அதிக எண்ணிக்கையிலான சொற்களை முதலில் பிரித்து அளிப்பதற்கு இந்த முறை பின்பற்றப்படுகிறது.



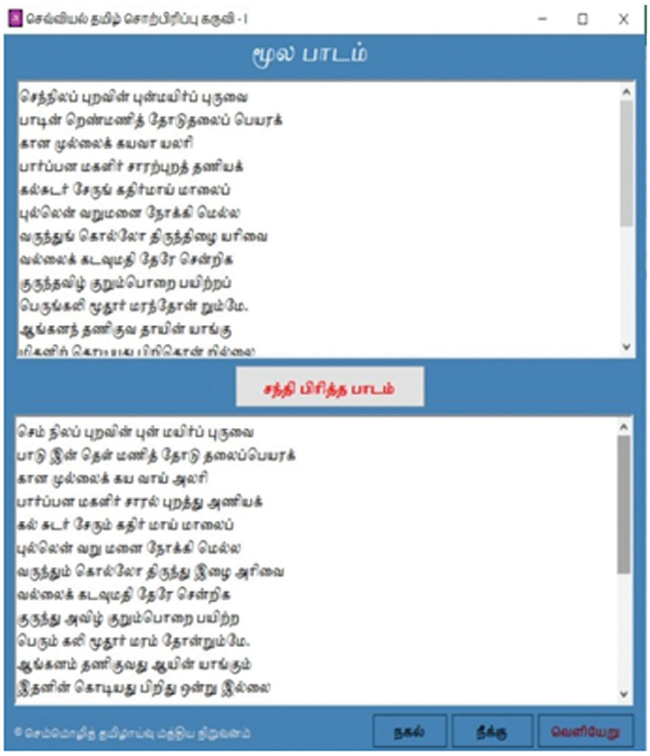
படம் 3: தரவு வரைபடம்

மேற்கண்ட தரவு வரைபடம் கருவி உருவாக்கத்தின் பல்வேறு படிநிலைகளை விளக்குகிறது. சொற்பிரிக்கப்பட வேண்டிய தரவுகள் உள்ளீடாகப் பெறப்படுகின்றன. பின்னர் ஏற்கெனவே உருவாக்கப்பட்ட உயர் அதிர்வெண் தரவுகளைக் கொண்ட அகராதியில் சென்று அச்சொல்லைத் தேடுகிறது. செவ்வியல் நூல்கள் 41இல் மூலப்பாடத்தில் இடம்பெற்றுள்ள மொத்தச் சொற்கள் 2,28,098. இந்தச் சொற்களில் அதிக எண்ணிக்கையிலான சொற்களைக் கணினி முதலில் மெட்டா தரவின் அடிப்படையில் தேடி எடுத்துக்கொள்ளும். இந்நிரலாக்கத்தின் முன் செயலாக்கம் மற்றும் முக்கியச் சொற்களின் தேர்வு என்பது அடிப்படையானது. உள்ளீடு செய்வதற்கான நிரல்கள் உருவாக்கப்பட்ட பின்னர் தரவைச் செயலாக்கி, ஒற்றை அல்லது கூட்டுச் சொற்களின் வடிவத்தில் முதன்மைச் சொற்களைப் பிரித்தெடுப்பதற்கான விதிமுறைகளை நிரல்கள் உருவாக்குகின்றன. அதன் பின்னர் இயல்பூக்கம் விதிப்படி தேவையற்ற இடைவெளிகள் போன்றவற்றை நீக்குகிறது. வார்த்தை உட்பொதிப்பின் அடிப்படையில், சொற்களின் தொடர்பைக் கணக்கிடுவதற்கான விதிகள் கணினிக்கு அளிக்கப்படுகின்றன. தேடி அச்சொல் கிடைத்தவுடன் அதை முக்கியச் சொல்லாக மெட்டா தரவில் சேமிக்கிறது. பின்னர் கணினிக்கு அளித்துள்ள பிரிப்பு விதிமுறைகளுக்கு உட்படுத்தி அச்சொற்களைப் பிரித்து அளிக்கிறது. இச்சுழற்சி முறை உள்ளீடு செய்யப்பெற்ற தரவின் அனைத்துச் சொற்களும் பிரித்து அளிக்கும் வரை நடைபெற்று இறுதியாக நிறைவுபெற்று முடிவுகளை அறிவிக்கிறது.

6. சொற்சூழல் கருவி அமைப்பு முறை

6.1 சொற்சூழல் கருவி 1.0

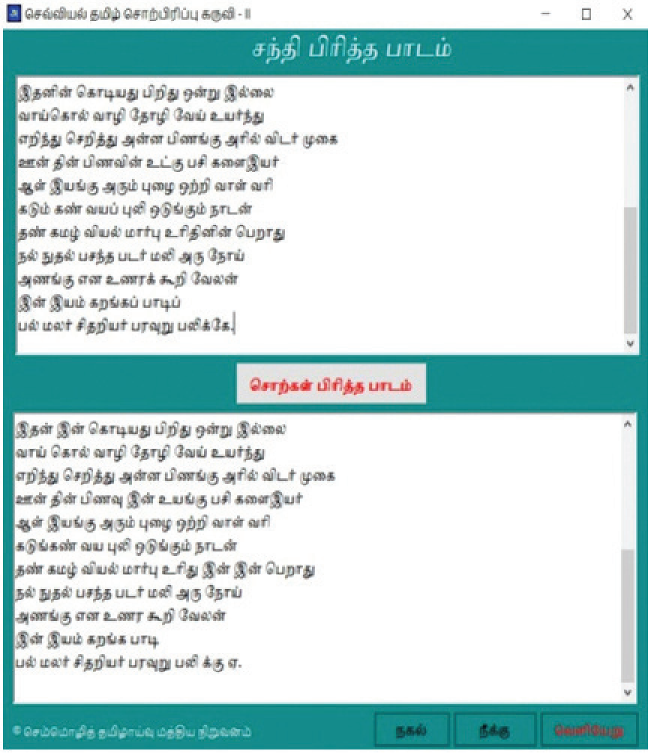
சொற்சூழல் கருவி மூலப்பாடத்திலிருந்து சந்தி பிரித்த பாடத்தைப் பிரித்து அளிக்கும் கருவியாக உருவாக்கப்பட்டுள்ளது. இதில் மூலப்பாடத்தில் இடம்பெற்றுள்ள பாடல் அல்லது நூற்பாக்களை உள்ளீடாகப் பின்வரும் படத்தில் உள்ள உரைப் பெட்டியில் அளித்துப் பின்னர் ‘சந்தி பிரித்த பாடம்’ என்னும் குமிழைச் சொடுக்குவதன் மூலம் உரைப்பெட்டியில் சந்தி பிரித்த பாடத்தைப் பெறலாம்.



படம் 4: சொற்சூழல் கருவி 1. 0

6.2 சொற்சூழல் கருவி 2.0

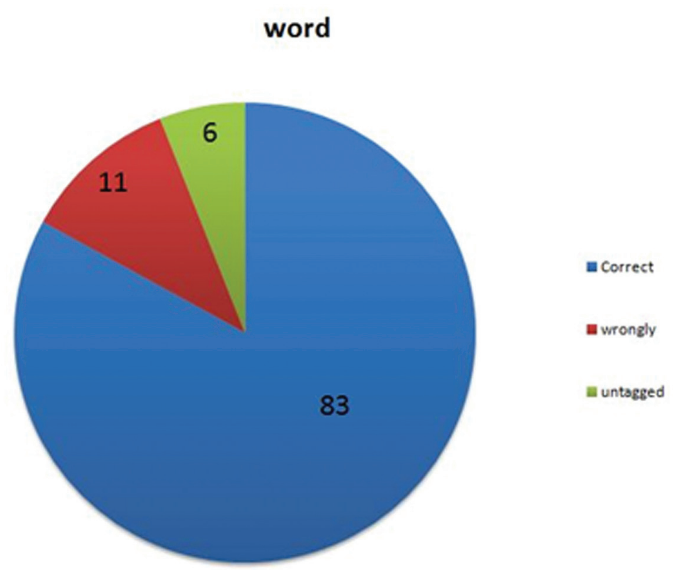
சந்தி பிரித்த பாடத்திலிருந்து சொற்கள் பிரித்த பாடத்தைப் பிரித்து அளிக்கும் கருவியாகச் சொற்சூழல் கருவி 2.0 உருவாக்கப்பட்டுள்ளது. இதில் சந்தி பிரித்த பாடத்தில் இடம்பெற்றுள்ள பாடல் அல்லது நூற்பாக்களை உள்ளீடாகப் பின்வரும் படத்தில் உள்ள உரைப்பெட்டியில் அளித்துப் பின்னர் ‘சொற்கள் பிரித்த பாடம்’ என்னும் குமிழைச் சொடுக்குவதன் மூலம் உரைப்பெட்டியில் சொற்கள் பிரித்த பாடத்தைப் பெறலாம்.



படம் 5: சொற்சூழல் கருவி 2. 0

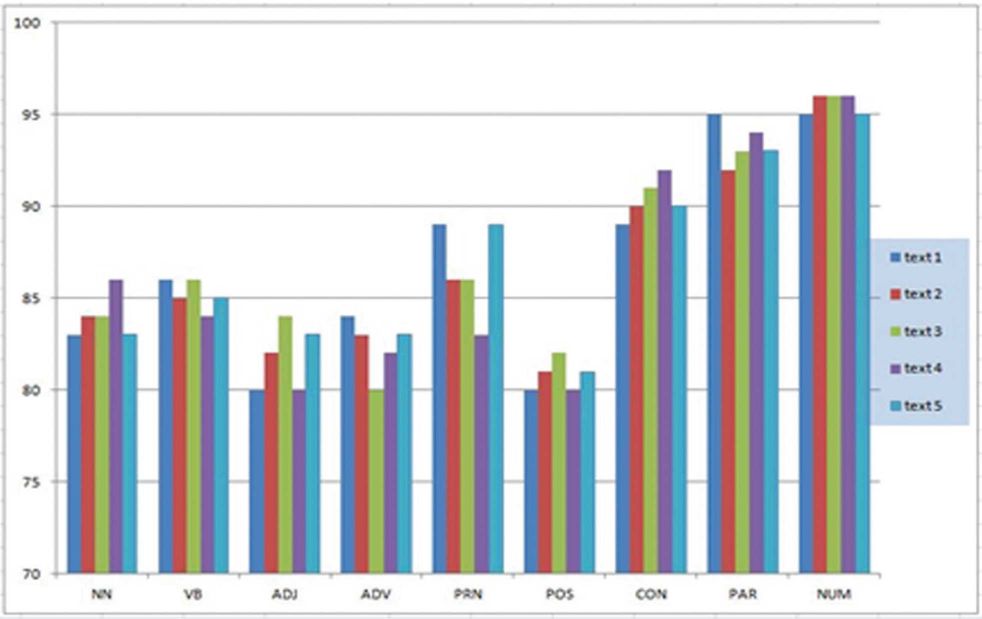
7. சிக்கல்கள்

சொற்சூழல் கருவி உருவாக்கத்தைப் பொறுத்தவரையில் உருவாக்கத்தின்போது இரண்டு வகையான சிக்கல்கள் உள்ளன. ஒன்று தரவுவழிச் சிக்கல் செவ்வியல் தரவுகளில் இடம்பெற்றுள்ள சொற்கள் வெவ்வேறு சூழல்களில் வெவ்வேறு விதமாகச் செயல்படுகின்றன. சில தொடர்களில், வார்த்தை மற்றும் தொடரியல் அமைப்பு ஒரே மாதிரியாக இருக்கும். இலக்கண வகை வார்த்தையின் தொடரியல் கட்டமைப்பைப் பொறுத்தது. சில பாடல்களில் வார்த்தை வடிவம் மற்றும் தொடரியல் அமைப்பு ஒரே மாதிரியாக இருக்கும். இந்த வகையான தரவுகள் நிலைகள் மிகவும் சிக்கலானவை. இயந்திரம் அத்தகைய சொற்களைத் தவறாகப் பிரித்து அளித்துவிடும்.



படம் 6: சொற்பிரிப்பு முடிவுகள்

நிரலாக்கச் சிக்கலைப் பொறுத்தவரையில் கருவி தவறாகப் பிரித்த சொற்களை இனம்கண்டு அவற்றை விதிகளின் அடிப்படையில் புதிய விதிகளை நிரலாக்கம் செய்து கணினிக்குத் தொடர்ந்து அளித்துக்கொண்டே இருக்க வேண்டும். அவ்வாறு தொடர்ந்து விதிகளை அளிக்கும் போது ஏற்கெனவே உள்ள தரவுகளின் முடிவுகளை இப்புதிய விதிகள் மாற்றம் செய்ய வாய்ப்பு உள்ளது. எனவே அத்தரவுகளைத் தொடர்ந்து கண்காணித்துக்கொண்டே இருக்க வேண்டும். மேற்கண்ட தரவுகளின் அடிப்படையில் உருவாக்கப்பட்ட கருவி 83% முடிவுகளைச் சரியாக அளித்துள்ளது.



படம் 7: இலக்கண வகை முடிவுகள்

இந்த முடிவுகளை இலக்கண வகைகளின்படி ஆய்விற்கு உட்படுத்தும்போது பின்னுருபுச் சொற்கள் அதிகச் சிக்கல்களை ஏற்படுத்துவதாக அமைகிறது.

8. முடிவுரை

இந்தக் கருவி செவ்வியல் நூல்களுக்கான சொற்பிரிப்புக்காக உருவாக்கப்பட்டது. இதற்கான தரவுகள் செம்மொழித் தமிழாய்வு மத்திய நிறுவனத்தின் உ.வே.சா செம்மொழித் தரவகத்திலிருந்து எடுத்துப் பயன்படுத்தப்பட்டுள்ளது. இக்கட்டுரையின் நோக்கம் செம்மொழி தமிழ் இலக்கிய ஆய்வை எளிதாக்கும் மொழி ஆய்வுக் கருவிகளை உருவாக்குவதற்கான தரவகங்களையும் அதற்கான கருவிகளையும் வடிவமைப்பது அடிப்படையாக அமைகிறது. செவ்வியல் தமிழ்த் தரவுகள் பாடல் வடிவிலும் நூற்பா வடிவிலும் உள்ள அமைப்பைக் கொண்டுள்ளது. எனவே செவ்வியல் ஆய்வுரையை எளிதாக்க உதவும் வகையில் இதுபோன்ற கருவிகளை உருவாக்க வேண்டிய அவசியம் உள்ளது. இதன் வழியாகக் கிடைக்கும் தரவுகள் இயற்கை மொழி ஆய்விற்கு அடிப்படை ஆதாரமாக அமையும். மேலும் ஏற்கெனவே உருவாக்கப்பட்டுள்ள இலக்கண விதிகள் திருத்தப்பட்டுச் செம்மைப்படுத்தப்படவும் இது பயன்படும். இவற்றைச் செம்மைப்படுத்துவற்கு மொழி வல்லுநர்கள், தமிழறிஞர்கள் துணைகொண்டு சொற்களைப் பகுப்பாய்வு செய்வதற்கான விதிகளை உருவாக்க வேண்டும். மேலும் விதிவிலக்கான தரவுகளுக்குப் புதிய விதிகளை உருவாக்கிக் கணினிக்கு அளிக்க வேண்டும். இவை அனைத்தும் முழுமை அடைந்து மேலும் தரவுகளை இணைக்கும்போது இக்கருவியை தற்காலத் தமிழ் மற்றும் பக்தி இலக்கிய ஆய்வுகளுக்கும் பயன்படுத்த முடியும்.

துணைநூல்கள்

1. Akshar Bharati, Vineet Chaitanya and Rajeev Sangal. 1995. "Natural language Processing A Paninian Perspective", Prentice Hall of India, India,
2. Dipanjan Das, Andre F.T., and Martins. 2011. "A Survey on Automatic Text Summarization", Language Technologies Institute, Carnegie Mellon University, 2007
3. Ela Kumar, "Natural Language Processing", I.K. International Publishing House Pvt. Ltd., New Delhi,
4. Gupta, V., and Lehal, G.S. 2009. "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Volume 1: 1, pp. 60-76,
5. Mathiyan, P. 2007. "canka ilakkiyac collataivu", Tamil University, Thanjavur
6. Nisheeth Joshi, and Iti Mathur. 2012. "Evaluation of Computational Grammar Formalisms for Indian Languages", International Conference in Computer Engineering and Technology, organized by Jodhpur Institute of Engineering and Technology, Jodhpur,
7. Rajam, V.S. 1992. "A Reference Grammar of Classical Tamil Poetry", The American Philosophical Society
8. Akilan R., Naganathan E R. 2015. "Morphological Analyzer for Classical Tamil texts a rule-based approach", ARPN Journal of Engineering and Applied Sciences, Volume : 10, No: 20, ISSN : 1819-6608
9. Ramaswami, N. 2001. "Lexical Formatives and Word Formation Rules in Tamil", in the journal of Languages in India, Volume 1: 8
10. https://ctlt.cict.in/
11. https://cict.in/

தமிழ்ச் சொல்வலை உருவாக்கமும் சவால்களும்

இராசேந்திரன் சங்கரவேலாயுதன்

ஆய்வுச்சுருக்கம்

இங்கு தமிழ்ச் சொல்வலை உருவாக்கம் பற்றி விளக்கப்படுகின்றது. இந்திய அரசின் நிதி நல்கையில் திராவிட மொழிகளின் சொல்வலைகள் திட்டத்தில் உருவாக்கப்பட்ட தமிழ்ச் சொல்வலை இந்தி சொல்வலை அடிப்படையில் அமைகின்றது. இத்தமிழ்ச் சொல்வலை கிட்டத்தட்ட 25,000 ஒருபொருள்பன்மொழியக் குழுமங்களைக் (synsets) கொண்டுள்ளது. இது தமிழின் மொத்தச் சொற்றொகையையும் கணக்கிடும் போது மிகக்குறைவான உள்ளடக்கம் ஆகும். தமிழ்ச் சொற்றொகையில் ஐந்து சதவீதம் கூட இந்தி-தமிழ் சொல்வலையில் உட்படுத்தப்படவில்லை என்று கூற இயலும். இனி இணைக்கப்படவேண்டிய கருத்துருக்களும் அவை உருப்படுத்தம் செய்யும் சொற்களும் ஏராளம்.

இராசேந்திரன் சங்கரவேலாயுதன்
மின்னஞ்சல்: rajushush@gmail.com

1. அறிமுகம்

சொல்வலை என்பது சொற்களைக் கருத்துருக்களாகக் கருதி அவற்றை ஒருபொருள் பன்மொழியக் குழுமங்களாகப் (synset) பகுத்து அவைகளுக்கிடையே உள்ள பொருண்மை மற்றும் சொல் உறவுகளை நிறுவி ஒரு வலையமைப்பாக வடிவமைப்பதாகும். சொல்வலை என்பதை சுருக்கமாக விளக்கவேண்டுமானால் இது ஒரு அகராதிப் பண்புக்கூறுகளும் பொருட்புல அகராதிப் (சொற்களஞ்சியம்) பண்புக்கூறுகளும் கொண்ட ஒரு சொல் மூலவளமாகும். தமிழுக்கு "காட்சி மூலப்பொருண்மையியல் சொற்களஞ்சியம்" உருவாக்கப்பட்டுள்ளது. இச்சொற்களஞ்சியம் தமிழ்ச் சொல்வலையாக விரிவுபடுத்தவேண்டும். சொற்கள் பெயர், வினை, பெயரடை, வினையடை போன்ற சொற்பாகுப்பாடுகளின் அடிப்படையில் பிரிக்கப்படாமல் நைடா முன்மொழிவதுபோல் பருப்பொருள்கள், நிகழ்வுகள், அருவங்கள், உறவன்கள் எனப் பாகுபடுத்தப்பட்டு பின்னர் சொற்பாகுபாடுகளுடன் இணைப்பது மொழியைத்தாண்டிச் சொற்களை இணைப்பது எளிமையானதும் மேம்பட்டதும் ஆகும். பயன்படுத்துவோருக்கு எளிமையான முறையில் சொல் குறித்த தேவையான செய்திகளைப் பெறவேண்டி கணினிமயமாக்கப்பட்டு ஒரு இணையதளச் சொல் தரவுதளமாக இதை எளிதில் மேம்படுத்தலாம்.

இந்தோ சொல்வலை

ஐ.ஐ.டி பம்பாயில் (IIT Bombay) உள்ள கணினி அறிவியல் மற்றும் பொறியியல் துறையில் இந்திய மொழி தொழில்நுட்ப மையத்தில் (Center for Indian Language Technology CFILT/சி.எஃப்.ஐ.எல்.டி) இயற்கை மொழி செயலாக்க குழுவால் இந்தி சொல்வலை 2000-ஆம் ஆண்டில் உருவாக்கப்பட்டது.

இந்திய மொழிச் சொற்களை உருவாக்கும் பெரிய நாடு தழுவிய திட்டம் இந்தோ சொல்வலைத் திட்டம் என்று அழைக்கப்பட்டது. இந்தோ சொல்வலை என்பது இந்தியாவின் 18 திட்டமிடப்பட்ட மொழிகளின் சொல்வலைகளால் இணைக்கப்பட்ட சொல்சார் அறிவுத் தளமாகும். இந்தி சொல்வலையிலிருந்து விரிவாக்க அணுகுமுறையைப் பயன்படுத்தி சொல்வலைகள் உருவாக்கப்பட்டன. இந்தி சொல்வலை இந்திய மொழியின்

முதல் சொல்வலையாகும். இது ஏற்றுக்கொள்ளப்பட்ட முறை ஆங்கிலத்திற்கான பிரின்ஸ்டன் சொல்வலை போலவே இருந்தது.

இந்தோ சொல்வலை பின்பற்றிய மூலோபாயத்தின் அடிப்படையில் போலந்து சொல்வலை, பிரின்ஸ்டன் சொல்வலையுடன் பொருத்தப்பட்டது/இணைக்கப்பட்டது.

திராவிட மொழிகளின் சொல்வலையாக்கம்

தமிழ் சொல்வலை உருவாக்க நடவடிக்கைகள் 2000-களிலேயே தொடங்கப்பட்டுவிட்டன. சென்னை ஏ.யூ.-கே.பி.சி ஆராய்ச்சி மையத்தில் தஞ்சைத் தமிழ்ப் பல்கலைக்கழகத்துடன் இணைந்து தமிழ் இணையப் பல்கலைக்கழகத்தின் (இன்றைய தமிழ் இணையக் கல்விக்கழம்) நிதி நல்கையுடன் (4 லட்சம்) தமிழ்ச் சொல்வலை உருவாக்கப் பணி தொடங்கப்பட்டது. இராஜேந்திரனின் (2001) தமிழ்ச் சொற்களஞ்சியத்தின் மூலப்பொருண்மையியல் கட்டமைப்பைப் பயன்படுத்தி ஒரு சொல்வலை அமைப்பு உருவாக்கப்பட்டு நிதி நல்கை நிறுவனத்திற்குத் தரப்பட்டது.

சென்னையில் நடைபெற்ற ஒரு பட்டறையின் போது திராவிடச் சொல்வலையின் உருவாக்கம் பற்றி விவாதிக்கப்பட்டு ஒரு திட்டவரைவு தயாரிக்கப்பட்டது. "ஆங்கிலத்தை திராவிட மொழிகளில் மொழிபெயர்ப்பதற்கான இயந்திர மொழிபெயர்ப்பு கருவிகளை உருவாக்குதல்" ("Developing Machine Translation tools for translating English into Dravidian Languages") என்ற தலைப்பில் ஒரு திட்டம் மனிதவள அமைச்சகத்தால் 2009இல் அனுமதிக்கப்பட்டது. இதில் திராவிட மொழிகளுக்கான உருவாக்கம் ஒரு பாகமாக அமைந்தது. திராவிட மொழிகளின் சொல்வலை உருவாக்கத்தில் நான்கு நிறுவனங்கள் பங்கேற்றன. இந்த திட்டம் 2011இல் முடிவுக்கு வந்தது. கிட்டத்தட்ட 15000 ஒருபொருள்பன்மொழிகளை தங்களுக்கு ஒதுக்கப்பட்ட மொழிகளுக்காக ஒவ்வொரு நிறுவனமும் முடித்தளித்தது.

மீண்டும் திராவிட மொழிகளின் சொல்வலைகளின் உருவாக்கம் இந்திய அரசின் மின்னணு மற்றும் தகவல்தொடர்புத் தொழில்நுட்பத் துறையின் நிதி நல்கையில் "திராவிட சொல்வலைகளின் உருவாக்கம், "தெலுங்கு, தமிழ், கன்னடம் மற்றும் மலையாளத்திற்கான ஒருங்கிணைந்த சொல்வலை (Development of Dravidian WorldNet: An Integrated WordNet for Telugu, Tamil, Kannada and Malayalam)" என்ற திட்டத்தின் கீழ் 2011 திசம்பரில் தொடங்கப்பட்டது. திராவிட மொழிகளின் சொல்வலைகளின் உருவாக்கம் இந்தோ சொல்வலையின் ஒருபாகமாக அமைந்தது. திராவிட சொல்வலையின் உருவாக்கச் செயல்பாடு மைசூர் பல்கலைக்கழகத்தில் கன்னடச் சொல்வலை, தமிழ் பல்கலைக்கழகத்தில் தமிழ்ச் சொல்வலை, திராவிடப் பல்கலைக்கழகத்தில் தெலுங்குச் சொல்வலை, அமிர்தா விஷ்வ வித்யபீடம் பல்கலைக்கழகத்தில் மலையாளச் சொல்வலை என மேற்கொள்ளப்பட்டது. 35000 ஒருபொருள்பன்மொழிகளை இலக்காக்கி இத்திட்டம் 2015 வரை நடைபெற்றது.

மனிதவள மேம்பாட்டு அமைச்சின் நிதி நல்கையில் தஞ்சையிலுள்ள தமிழ்ப்பல்கலைக் கழகமும் கோயம்புத்தூரில் உள்ள அமிர்தா விஷ்வ வித்யாபீடமும் குப்பத்திலுள்ள திராவிடப் பல்கலைக் கழகமும் மைசூரிலுள்ள மைசூர் பலகலைக்கழகமும் இணைந்து திராவிட மொழிகளின் சொல்வலைகளை உருவாக்கும் திட்டதைச் செயல்படுத்தின. இச்சொல்வலைகள் இந்தோ சொல்வலை உருவாக்கத் திட்டத்தின் பகுதியாக அமைந்து. திராவிட மொழிகளின் சொல்வலைகள் ஏற்கனவே உருவாக்கப்பட்டு இணையதளத்தில் பயன்பாட்டில் விடப்பட்டுள்ள இந்தி சொல்வலையை அடிப்படையாகக்கொண்டு விரிவாக்க முறை அடிப்படையில் திராவிட மொழிகளின் சொல்வலைகள் உருவாக்கப்பட்டன. தமிழ்ப் பல்கலைக்கழகம் தமிழ்ச் சொல்வலையையும் திராவிடப் பல்கலைக்கழகம் தெலுங்கு சொல்வலையையும் அமிர்தா பல்கலைக்கழகம் மலையாளச் சொல்வலையையும் மைசூர் பல்கலைக்கழகம் கன்னடச் சொல்வலையையும் உருவாக்கி வருகின்றன. இந்தத் திட்டத்தின் முக்கியக் குறிகோள்கள் பின்வருவனவாகும்:

- திராவிட மொழிகளுக்கு விரிவான உயர்ந்த சிறப்புவாய்ந்த பன்மொழியச் சொல்சார் தரவுமையம் உருவாக்குதல்
- மொழிச் சொல்வலைகளை ஒன்றாக இணைக்கும் மொழிச் சுதந்திரமான பொருண்மைக் கருத்துருக்களை உருவாக்குதல்
- எல்லாத் திராவிட மொழிகளுக்கும் தகவலின் பொருண்மையியல் பாகுபாட்டை நிலைபேறாக்கம் செய்தல் மற்றும் பயன்பாடுகளின் உருவாக்கத்திற்கு மூலவளங்களைத் தருதல்
- ஏற்கனவே இருக்கும் மூலவளங்களைப் பயன்படுத்திச் சொல்வலைகளை உருவாக்குதல்

2. தமிழ்ச் சொல்வலைக்கான மூலவளங்கள்

இந்தோ சொல்வலையின் பாகமாக அமையும் தமிழ்ச் சொல்வலை ஆங்கிலச் சொல்வலையுடன் இந்தி சொல்வலை வழி இணைக்கப்பட்டுள்ளது. தமிழ்ச் சொல்வலை இந்தி சொல்வலையின் அடிப்படையில் அமையாமல் தனியாக உருவாக்கப்பட்டால் சிறப்பாக அமையும் என்று கூற இயலும். இத்தகைய தமிழ் சொல்வலை உருவாக்கத்திற்கு மூல வளங்களாகத் தமிழ்ப் பல்கலைகழக மொழியியல் துறையில் இராசேந்திரனின் மேற்பார்வையில் சமர்ப்பிக்கப்பட்ட

ஆய்வேடுகளும் இராசேந்திரனால் செய்யப்பட்ட தமிழ்ச் சொற்பொருண்மை ஆய்வுகளும் (சக்திவேல் & இராசேந்திரன் 1994, இராசேந்திரன் 2001, இராசேந்திரன் & பாஸ்கரன் 2006, இராசேந்திரன் & பாக்கியராஜ் 2019) அமையும். இராசேந்திரன் (2001) தற்காலத் தமிழ்ச் சொற்களஞ்சியம் இச்சொல்வலையாக்கத்திற்குப் பெரிதும் உதவும். அவரது சொற்களஞ்சியத்தின் மூலம் வெளிப்படும் மூலப்பொருண்மையியலும் (ontology) அவர்தம் மூலப்பொருண்மையியல் ஆய்வுகளும் (இராசேந்திரன் & அனிதா 2019) தமிழ்ச் சொல்வலை உருவாக்கத்திற்கு பெரிதும் பயனுள்ளதாக அமையும். மேலும் தமிழுக்கென்ற சொல்வலை உருவாக்கத்தில் முன்னோடியாக விளங்கும் தொடக்கநிலையில் உள்ள தமிழ்ச் சொல்வலை அவரால் உருவாக்கப்பட்டு இணையதளத்தில் திறந்த மூலமாக (open source) இடப்பட்டுள்ளது. இதுவும் தமிழ்சொல்வலை உருவாக்கத்திற்கு துணையாக அமையும்.

3. அகராதியும் சொல்வலையும்

அகராதியானது சொல்லின் எழுத்துவடிவு, பொருள், பயன்பாடு இவற்றைக் காணப் பொதுவாய்ப் பயன்படுத்தப்படுகின்றது. ஆனால் சொல்வலை ஒரு பொருண்மை அகராதியாகும் (semantic dictionary). சொல்வலை சொற்களுக்கு வர்ணனையும் விளக்கமும் மாதிரி எடுத்துக்காட்டுகளும் தரும். சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளை மிகத்தெளிவாகப் பயன்படுத்தக்க விதத்தில் கூறும். உச்சரிப்பு, சொல்லாக்க உருபனியல் செய்திகள், சொல்மூலம் அல்லது வரலாறு இவற்றைத் தராது.

4. சொற்களஞ்சியமும் சொல்வலையும்

முதல் சொற்களஞ்சியம்/பொருட்புல அகராதி (thesaurus) ராஜெஸ்ட் என்பரால் உருவாக்கப்பட்டது.

இது பொருண்மை ஒழுங்கமைப்பைக் கொள்கையாகக்கொண்டு உருவாக்கப்பட்டது. கருத்துரு அடிப்படையில் அமைக்கப்பட்டது. பயன்பாட்டாளர் தங்கள் மனதிலுள்ள கருத்துருவைக் கொண்டு சொற்களைக் கண்டுகொள்ள இது உதவும். சொல்வலைகளும் சொற்களஞ்சியம் போல அமையும். இதன் கட்டுமான அலகு ஒருபொருள்பன்மொழியக் குழுமம் ஆகும். ஒருபொருள்பன்மொழியக் குழுமத்தில் ஒரு கருத்துருவை வெளிப்படுத்தும் எல்லாச் சொற்களும் குழுமம் செய்யப்பட்டிருக்கும். சொல்வலையைப் பயன்படுத்துபவர் மனதிலுள்ள கருத்துருவைக் கொண்டு சொற்களைக் கண்டுகொள்ள இயலும். ஒருபொருள்பன்மொழியக் குழுமங்கள் உள்ளடங்கு–உள்ளடக்கு மொழியம் (hyponymy–hypernymy), சினைமொழியம்–முழுமொழியம் (meronymy–holonymy or part and hole), உட்படுத்து மொழியம் (entailment) போன்ற உறவுகளால் இணைக்கப்பட்டுள்ளன. சொல்வலை கருத்து நிலையையும் சொல் நிலையையும் தெளிவாகப் பிரிக்கின்றது. இந்த வேறுபாடு சொல் உறவுகளுக்கும் கருத்துருக்களுக்கும் உள்ள வேறுபாட்டில் பிரதிபலிக்கின்றது. சொற்களஞ்சியத்திற்கு அப்பாற்பட்டு சொல்வலையில் சொற்களுக்கும் கருத்துருக்களுக்கும் உள்ள உறவுகள் வெளிப்படையாகக் கூறப்பட்டு அடையாளப்படுத்தப்பட்டுள்ளன. பயன்படுத்துபவர் உறவுகளைத் தேர்ந்தெடுத்து, ஒரு கருத்துருலிருந்து மற்றொரு கருத்துருவுக்குச் சென்று கருத்துரு வெளியிடத்தில் வலம் வர இயலும்.

5. சொல் வலையில் உறவுகள்

சொல்வலை கருத்துரு–பொருண்மை உறவுகளுக்கும் சொற்களை இணைக்கின்ற சொல் உறவுகளுக்கும் வேறுபாடு காட்டுகின்றது. ஆனால் சொல் நோக்கில் கவனம் செலுத்துபவர் முதன்மையாகச் சொல் உறவுகளைப் பயன்படுத்துவர். சொல்வலை பொருண்மை உறவுகளால் ஒழுங்கமைக்கப்பட்டுள்ளது. சொல்வலை வேறுபட்ட தொடரியல் வகைப்பாடுகளில் அடங்கும் சொற்களை இணைக்கும் அடுக்கு உறவுகளைக் கொண்டிருக்கவில்லை. நான்கு முக்கியத் தொடரியல் வகைப்பாடுகளான பெயர், வினை, பெயரடை, வினையடை என்பன தனித்தனியாகத் தரப்பட்டுள்ளன. பெயர்கள் படிநிலை அமைப்புகளாகவும் வினைகள் வேறுபட்ட உட்படுத்து உறவுகளாலும்; (entailment relation) பெயரடைகளும் வினையடைகளும் இ–பரிமாண உயர்வெளியிடத்திலும் ஒழுங்கமைக்கப்பட்டுள்ளன.

6. பெயர்ச்சொல்வலை

பெயர்ச்சொல்வலையில் அடிப்படை பொருண்மை உறவு ஒருபொருள்பன்மொழிய உறவாகும். ஒருபொருள்பன்மொழியக் குழுமங்கள் அடிப்படைக் கட்டுமான அலகுகளாகும். ஒருபொருள்பன்மொழியக் கருத்துச்சாயல் என்பது எல்லாச் சூழல்களிலும் ஒன்றையொன்று இடம்பெயர்த்தல் என்பதை உட்படுத்தாது. இடம்பெயர்த்தல் என்ற அளவியால் இயற்கை மொழிகளுக்குக் குறைந்த எண்ணிக்கையிலான ஒருபொருள் பன்மொழிகள் மட்டுமே வரும். சொல்வலையில் ஒரு பொருள் பன்மொழிகளைக் குறைந்தது சில சூழல்களில் இடம் பெயர்க்கலாம். ஒருபொருள்பன்மொழியக் குழுமத்திற்கு ஒரு பொருள்தான் உண்டு. எ.கா. {நூல், புத்தகம், புஸ்தகம்} “படிப்பதற்கு ஏற்றவகையில் அட்டைபோட்டு இணைத்த அச்சடித்த தாள்களின் தொகுப்பு“ சொல்வலையில் ஒருபொருள்பன்மொழியக் குழுமங்கள் பொருண்மை உறவுகளால் இணைக்கப்பட்டுள்ளன. பெயர்களை ஒழுங்குபடுத்தப் பயன்படுத்தப்படும் மிக முக்கியமான

உறவு உள்ளடங்கு மொழியமாகும் (hyponymy). இந்த உறவுதான் சொற்களைச் சொற்படிநிலை அமைப்பில் ஒழுங்குபடுத்துகின்றது.

6.1. சொற் படிநிலை அமைப்பு

உள்ளடங்கு உறவுகள் பற்றிய செய்திகள் மரபு அகராதிகளில் வரையறை விளக்கங்களாகத் தரப்பட்டிருக்கும்.

எ.கா.

குயில் – இனிய குரலுடைய கரிய நிறப்பறவை.

பறவை – இரு கால்களும் அலகுகளும் உடைய, உடலில் இருபக்கங்களிலும் இறகுகள் உள்ள பறப்பதற்கு ஏற்றவகையில் சிறகுகளும் கொண்ட விலங்கினம்.

விலங்கினம் – தானாகவே இயங்கும் உணர்வு உறுப்புகளும் செல்லுலோஸ் இல்லாத செல் சுவர்களும் உள்ள உயிரினம்.

உயிரினம் – உயிர் வாழ்கின்ற ஒன்று.

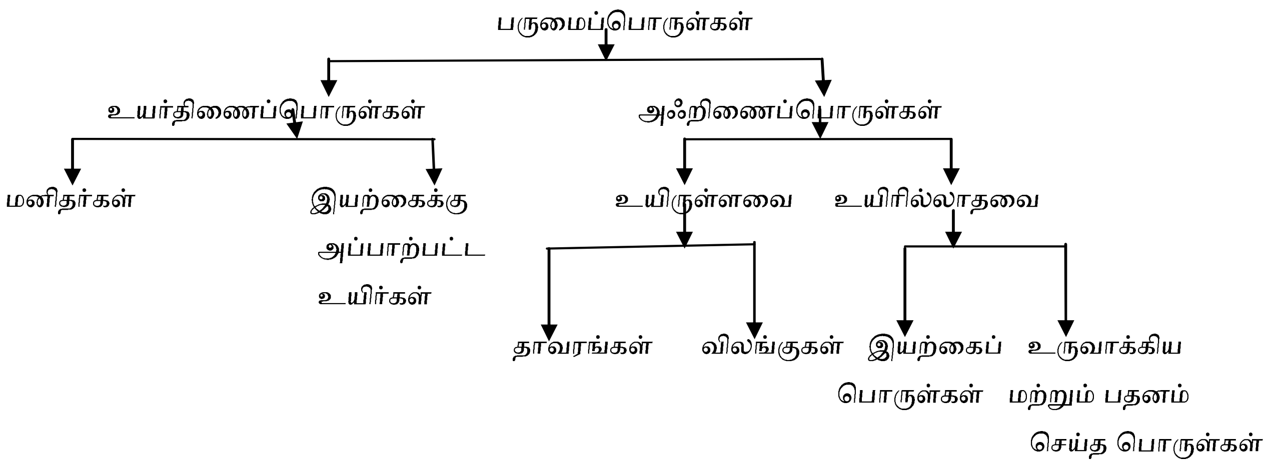
ஒவ்வொரு உள்ளடக்குச் சொல்லும் மிகப்பொதுவான உள்ளடக்குச் சொல்லுக்குக் கொண்டு செல்லும். உள்ளடக்கு மொழியத்தைச் சொற்களுக்கு இடையே உள்ள உறவாக உருப்படுத்தம் செய்ய இயலாது. உள்ளடக்கு மொழியம் அகராதிப்படுத்தப்பட்ட கருத்துருக்களுக்கு இடையே உள்ள உறவாகும். ஒரு சொற்படிநிலை அமைப்பை உள்ளடக்கு உறவால் இணைக்கப்பட்ட ஒருபொருள்பன்மொழியக் குழுமங்களின் தொடர்ச்சியால் மீட்டுருவாக்கம் செய்யலாம்.

{குயில்}@ → {பறவை}@ → {விலங்கினம்}@ → {உயிரினம்}

உள்ளடங்கு உறவுக்கு இணையான உள்ளடக்கு உறவையும் தரலாம். உள்ளடங்கு–உள்ளடக்கு மொழிய உறவுகளால் ஒரு படிநிலை ஒழுங்குமுறை வெளிப்படும். இந்த வகையில் வரும் படிநிலைகள் கணினியியலார்களால் அறிவை உருப்படுத்தம் (Knowledge representation) செய்யும் வழியாகப் பயன்படுத்தப்படுகின்றது.

6.2. தனித்தன்மையான தொடக்கிகள்

பெயர்கள் எல்லாவற்றையும் ஒரே படிநிலை அமைப்பில் தரவேண்டி படிநிலைக் கொள்கையை நீட்சி செய்யலாம். சொல்வலை பெயர்களைத் தனித் தன்மையான தொடக்கிகள் (unique beginners) அமையப் பல படிநிலை அமைப்புகளாகப் பகுத்துள்ளது. தனித்தன்மையான தொடக்கிகள் சொற்பொருண்மையியல் பொருண்மைக்கூறாய்வின் பொருண்மைக் கூறுகளுடன் ஓரளவுக்குப் பொருந்தும்.



மெய்ப்பொருள் மூலாய்வு அமைப்பைக் (ontological structure) காட்டும் வகையிலான சொற்களின் உருப்படுத்தம் சொற்களின் பொருண்மை மரபுரிமைச் செயற்பாங்கை (Lexical inheritance) வெளிக்கொணர்கின்றது.

எ.கா. {ஊர்தி, வாகனம்}@ → {நிலத்தில் ஓடும் ஊர்தி}@ → {பொறியால் ஓடும் ஊர்தி}@ → {நான்குசக்கர ஊர்தி}@ → {பயணஊர்தி}@ → {பேருந்து}}

List of 25 unique beginners for noun source files of EuroWordNet

{act, activity}	{natural object}
{animal, fauna}	{natural phenomenon}
{artifact}	{person, human being}
{attribute}	{plant, flora}
{body}	{possession}
{cognition, knowledge}	{process}
{communication}	{quantity, amount}
{event, happening}	{relation}
{feeling, emotion}	{shape}
{food}	{state}
{group, grouping}	{substance}
{location}	{time}
{motivation, motive}	

இராசேந்திரனால் உருவாக்கப்பட்ட தமிழ்ச் சொற்களஞ்சியம் (இராசேந்திரன் 2001) தமிழ்ப்பெயர்ச்சொற்களின் சொற்றொகையை நைடாவைப் பின்பற்றி (Nida 1976a) வகைப்பாட்டியல் வடிவில் தருகின்றது.

6.3. வேறுபடுத்தும் பொருண்மைக் கூறுகள்

பெயர்களின் படிநிலை அமைப்பு உள்ளடங்கு–உள்ளடக்கு உறவால் உருவாக்கப்பட்டாலும் விளக்கங்கள் ஒரு கருத்துருவை மற்றொரு கருத்துருவிலிருந்து பிரிக்கும் பொருண்மைக் கூறுகளால் தரப்படுகின்றது.

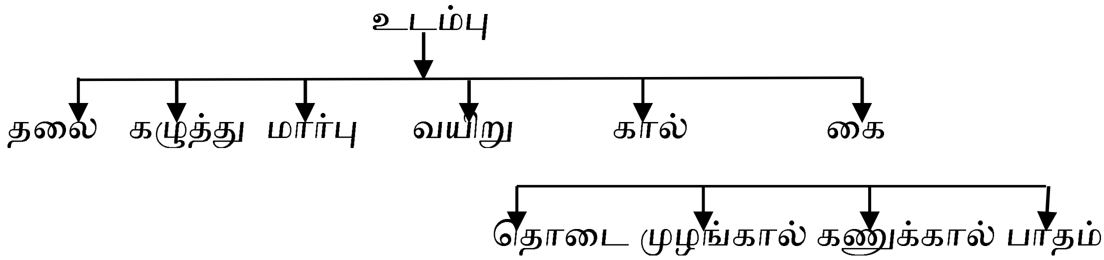
எ.கா. குயில் – பாடுகின்ற சிறிய கரிய நிறப்பறவை

குயில் என்பதை மூன்றுவகைப் பட்ட வேறுபடுத்தும் பண்புக்கூறுகளுடன் தொடர்புபடுத்தலாம்.

1. அடைகள்: சிறிய, கரிய
2. பாகங்கள்: அலகு, சிறிய
3. செயல்பாடுகள்: பாடு, பற

6.4. சினை–முழு மொழியம்

சொற்களைப் படிநிலை அமைப்பில் தருவதில் சினை–முழு உறவு முக்கியப் பங்கு வகிக்கின்றது.



சினை–முழு மொழிய உறவு பின்வரும் துணை வகைகளை உள்ளடக்கும்:

1. முழுப் பொருளுக்கும் அதன் உறுப்புகளுக்கும் இடையிலான உறவை வெளிப்படுத்துவன (பாகம், எ.கா. கை – விரல்)
2. முழுமைக்கும் அதிலிருந்து பிரிந்த பகுதிக்கும் உள்ள உறவை வெளிப்படுத்துவன (பகுதி, எ.கா. துண்டு – உலோகம்)
3. இடங்களுக்கும் விரிந்த இடங்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (இடம், எ.கா. பாலைவனச்சோலை – பாலைவனம்)
4. குழுமத்திற்கும் அதன் அங்கத்தினர்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (எ.கா. மந்தை – ஆடு)
5. பொருள்களுக்கும் அது உருவாக்கப்பட்ட உருப்பொருள்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (ஆன, எ.கா. புத்தகம் – காகிதம்)

பின்வரும் அட்டவணையில் பெயர்வலையில் கூறப்பட்டுள்ள பொருண்மை–சொல் உறவுகள் பட்டியலிடப்பட்டுள்ளன:

உறவுகள்	துணை வகைகள்	எடுத்துக்காட்டுகள்
ஒருபொருள் பன்மொழியம்		புத்தகம், நூல்
உள்ளடங்கு – உள்ளடக்கு மொழியம்		விலங்கு–பாலூட்டி
உள்ளடக்கு–உள்ளடங்கு மொழியம்		பசு–பாலூட்டி
முழு–சினை மொழியம்	முழுமை–பாகங்கள்	மேசை–கால்
	பாகம்–முழுமை	சக்கரம்–வண்டி
	அங்கத்தினர்கள்–குழு	படைத்தலைவர்–படை
	குழு–அங்கத்தினர்கள்	துறை–பேராசிரியர்
	பகுதி–முழுமை	துளி–கண்ணீர்
	இடம்–பரந்த இடம்	பாலைவனச்சோலை–பாலைவனம்
	பொருள்–உருப்பொருள்	புத்தகம்–தாள்
ஈரிணை எதிர்நிலை	நிரலாக்கம் செய்யத்தக்கது	நல்லவன்–கெட்டவன்
	துணை எதிர்நிலை	பகல்–இரவு
	தனிப்பட்டவை	அஃறிணை–உயர்திணை
	துருவ எதிர்நிலை	ஆண்–பெண்
	பரஸ்பர சமூகப்பாத்திரங்கள்	மருத்துவர்–நோயாளி
	சொந்தங்கள்	அம்மா–மகள்
	கால உறவுகள்	காலை–மாலை
	இட உறவுகள்: செங்கோண எதிர்நிலை	வடக்கு–கிழக்கு வடக்கு–மேற்கு மேற்கு–தெற்கு கிழக்கு–தெற்கு
	இட உறவுகள்: நேர் எதிர்நிலை	வடக்கு–தெற்கு
	கிழக்கு–மேற்கு	
	பல்லிணை எதிர்நிலை: சங்கிலி	ஒன்று, இரண்டு, மூன்று,...
	சுற்று	ஞாயிறு, திங்கள், செவ்வாய், புதன், வியாழன், வெள்ளி, சனி

7. வினைச் சொல்வலை

வினைகள் ஒரு மொழியின் சொல் மற்றும் தொடரியல் வகைப்பாட்டில் முக்கியப் பங்கு வகிக்கின்றது. இதன் பயனிலை–பங்கெடுப்பாளர் அமைப்பு (predicate argument structure) இது வரும் வாக்கியத்தின் சாத்தியமான அமைப்புகளைத் தீர்மானிக்கின்றது.

எ.கா.

காற்று வீசுகின்றது.

அவன் சென்னையிலிருந்து தஞ்சாவூர் வந்தான்.

அவள் மாம்பழம் சாப்பிடுகிறாள்.

அவர் அவளிடமிருந்து ஒரு லட்ச ரூபாய்க்கு ஒரு கார் வாங்கினார்.

7.1. வினைகளைப் பொருண்மைக் களங்களாகப் பிரித்தல்

யூரோ சொல்வலை வினைகளைச் 15 பொருண்மைக் களங்களாகப் பிரிக்கின்றது (Vossen 1998).

1. Verbs of bodily functions and care (Ex. sweat, shiver, faint, etc.)
2. Verbs of change (Ex. change, etc.)
3. Verbs of communication (Ex. stammer, appeal, bet, teach, creak, etc.)
4. Competition Verbs (Ex. fight, etc.)
5. Consumption Verbs (Ex. drink, etc.)
6. Contact Verbs (Ex. hit, scrub, wipe, etc.)
7. Cognition Verbs (Ex. infer, guess, assume, etc.)
8. Creation Verbs (Ex. engrave, print, etc.)
9. Motion Verbs (Ex. gallop, race, fly, swim, etc.)
10. Emotion or Psych Verbs (Ex. amuse, charm, etc.)
11. Stative Verbs (Ex. surround, cross, etc.)
12. Perception Verbs (Ex. watch, spy, etc.)
13. Verbs of Possession (Ex. have, rob, bestow, auction, etc.)
14. Verbs of Social Interaction (Ex. impeach, franchise, excommunicate, etc.)
15. Weather Verbs (Ex. rain, thunder, snow, hail, etc.)

இராசேந்திரன் (2001) நைடாவைப் (1975a) பின்பற்றி வினைகளைப் பௌதிக வினைகள் (பெய், வீசு), உடல்கூறு வினைகள் (பிரசவி, வியர்), புலனுணர்வு வினைகள் (கசு, குளிர்), உணர்ச்சி வினைகள் (கோபப்படு, வேதனையடை), அறிவுசார் வினைகள் (ஊகி, கணி), கருத்துப்பாற்ற வினைகள் (பேசு, இகழ்), சலன வினைகள் (நகர், நட), தர்க்க வினைகள் (அடி, உடை), உடைமைமாற்ற வினைகள் (கொடு, விடு), பல்கூட்ட வினைகள் (சமையல் செய்) எனப்பிரித்துள்ளார்.

7.2. வினைகளுக்குத் தனித்தன்மையான தொடக்கிகள்

வினைச் சொற்களஞ்சியத்தைப் பொருண்மைக் களங்களாகப் பிரிப்பது வினைத் தரவுகளை ஒழுங்கு முறைபடுத்துவதோடு சொற்களஞ்சியத்திலுள்ள எல்லா வினைகளுக்கும் ஒரு பொதுவான வேர் அல்லது தனித்தன்மையான தொடங்கிகள் இல்லாத குறையையும் நிவர்த்தி செய்கின்றது.

பெயர்களைப் போல வினைகளையும் ஒருபொருள் பன்மொழியக்குழுமங்களாகக் குழும இயலும். ஆனாலும் நாம் ஒருபொருள்பன்மொழி என்பது ஒன்றையொன்று இடம் பெயர்க்கும் சொற்கள் என்று வரையறை செய்தால் வினையில் ஒருபொருள்பன்மொழிகள் இல்லாமல் போகும் அல்லது மிக அரிதாகத் தான் இருக்கும்.

7.3. வினைகளின் பொருண்மைக் கூறாய்வு

வினைகளை அவற்றின் பொருண்மைப் பண்புக்கூறுகளால் வரையறை விளக்கம் செய்ய இயலும். வினைகளைச் சிறு பொருண்மைக்கூறுகளாகப் பிரிக்க இயலும் தன்மை காரணமாக எளிய செயல்களால் கலவைத் தன்மையான செயல்களை விளக்க இயலும்.

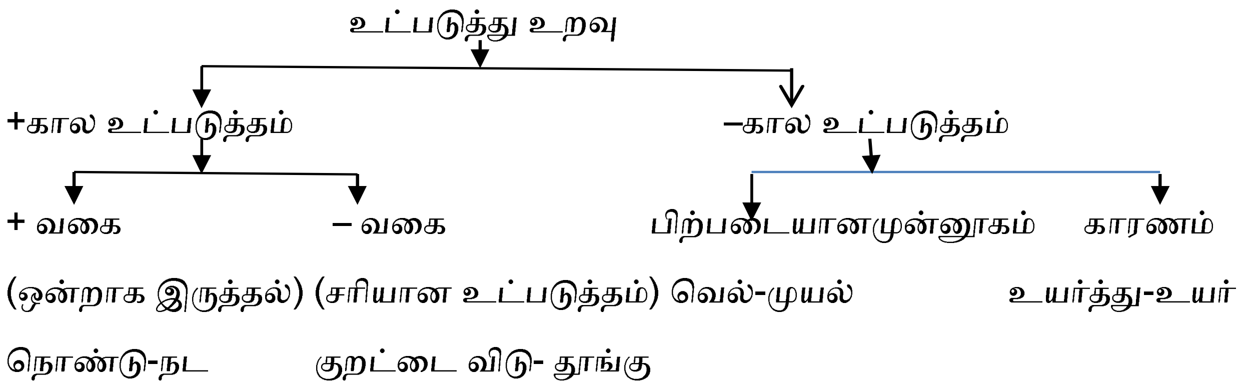
கொல் – சாகச் செய்தல்

எறி – ஒருவர் தன் கையிலிருந்தப் பொருளை விசையால் போகச் செய்தல்

ஓட்டு – ஒருவர் ஒன்றை ஓடச் செய்தல்

7.4. வினைகளுக்கிடையிலான சொல் மற்றும் பொருண்மை உறவுகள்

நாம் நான்கு வகையான உட்படுத்து உறவுகளை (entailment) வெளிப்படுத்தலாம்:



7.5. வினைகளுக்கிடையிலான பல்பொருள் ஒருமொழியம்

பெயர்களை விட வினைகள் எண்ணிக்கையில் குறைந்தவை. வினைகளின் பல்பொருள் ஒருமொழியம் பெயர்களின் பல்பொருள் ஒருமொழியத்தைக் காட்டிலும் அதிகம். வினைகளின் பொருண்மை நெகிழ்ச்சி அதன் பகுப்பாய்வைக் கடினப்படுத்துகின்றது. வினைகள் அவை எடுக்கும் பங்கெடுப்பவர் அமைப்பு (argument structure) அடிப்படையில் தங்கள் பொருளை மாற்றும். ஆனால் பெயர்களின் பொருண்மை வினைகளின் பொருண்மைகளைக் காட்டிலும் நிரந்தரமானது.

7.6. தொடரியல் பண்புகளும் பொருண்மை உறவுகளும்

தற்போது அகராதியில் சொற்களுக்கு அதன் தொடரியல் பண்புக்கூறுகளைத் தரும் போக்குக் காணப்படுகின்றது. வினைகளைப் பொருண்மை உறவுகளாக மட்டும் பார்ப்பது தொடரியல் பற்றிய சில செய்திகளைத் தரும். வினையைப்பற்றி தெரிந்து கொள்ள வினையின் தொடரியல் பண்புகளைச் சொல்வலையில் தருவது குறித்த சாத்தியம் ஆராயப்படவேண்டும்.

உறவுகள்	விளக்கம்/துணை வகை	எடுத்துக்காட்டு
ஒருபொருள் பன்மொழியம்	இடம்பெயர்க்க இயலும் செயல்கள்	தூங்கு–உறங்கு
பகுதி–முழுமை மொழியம்	உட்படும்–உட்படுத்தும் செயல்	பற–பிரயாணி
வகை உறவு	செயல்–துணைவகை	நட–நொண்டு
உட்படுத்து உறவு	செயல்–காரணச்செயல்	உயர்–உயர்த்து
”	செயல்–முன்னூகச்செயல்	வெல்–முயல்
எதிர்நிலை	எதிர்மறை	கூடு–குறை
	மறுதலை	வில்–வாங்கு
	திசை எதிர்நிலை	புறப்படு–வந்துசேர்

8. பெயரடை மற்றும் வினையடைச் சொல்வலை

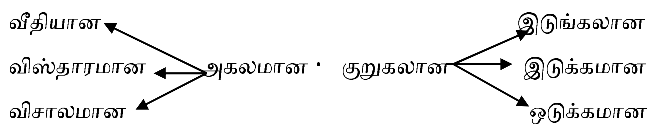
சொல்வலை பெயரடைகளை இரு முக்கியமான வகுப்புகளாகப் பிரிக்கின்றது: வர்ணனைப் பெயரடைகள் (descriptive adjectives) மற்றும் தொடர்புப் பெயரடைகள் (relational adjectives). வர்ணனைப் பெயரடைகள் தலைப் பெயர்களுக்கு இருதுருவ அடைகளின் மதிப்பீட்டைத் (values of bipolar attributes) தருகின்றது; எனவே ஈரிணை எதிர்நிலைகள் (binary oppositions) (எதிர்மறை– antonymy) மற்றும் பொருண்மை ஒற்றுமையால் (similarity of meaning) ஒழுங்கமைக்கப்படுகின்றது.

வர்ணனைப் பெயரடை (descriptive adjective): பெரிய, கனமான

தொடர்புப் பெயரடை (relational adjective): பொருளாதார, சகோதர

குறிப்பு அடைசெய்யும் பெயரடை (reference modifying adjectives): பழைய, முன்னாள்

ஒரு வர்ணனை அடை ஒரு பெயருக்கு ஒரு மதிப்பு தரும். எடுத்துக்காட்டாக, கனமான, இலேசான என்பது எடை என்பதன் அடைமொழிகளாகும்; தாழ்ந்த, உயர்ந்த என்பன உயரம் என்பதன் அடைமொழிகளாகும். எதிர்மறை (antonymy) தான் வர்ணனைப் பெயரடைகளில் அடிப்படையானதாகும். எதிர்மறை ஒருபொருள்பன்மொழியத்தைப் போன்று சொல்வடிவுகளுக்கு இடையிலான உறவாகும். எதிர்மறைப் பெயரடைகள் அடைமொழியின் எதிர்செய்யும் மதிப்புகளை வெளிப்படுத்துகின்றது (எ.கா. கனமான, இலேசான என்பன எடை அடைமொழியின் எதிர் துருவங்களாகும்.) நேரடியான எதிர்மறைகள் இல்லாத பெயரடைகள் (எதிர்மறை ஆற்றல் குறைவாய் உள்ள பெயரடைகள்) எதிர்மறைப் பெயரடைகளுடன் “போல” உறவு கொண்டிருக்கும்:



உறவுகள்	தொடர்புபடுத்தும் சொல்வகைப்பாடு	எடுத்துக்காட்டு
எதிர்மறை (நிரலாக்கம் செய்ய இயலுவது)	பெயரடை–பெயரடை	அழகான–குரூரமான
எதிர்மறை (நிரலாக்கம் செய்ய இயலாதது)	பெயரடை–பெயரடை	உயிருள்ள–செத்த
ஆக்கவடிவு	பெயரடை–பெயர்	அழகான–அழகு
அடைமொழி	பெயர்–பெயரடை	வடிவம்–சின்ன
தொடர்பு	பெயரடை–பெயர்	பொருளாதார–பொருளாதாரம்
போல	பெயரடை–பெயரடை	பாரமான–கனமான

9. சொல்வலையை திட்டமிட்டு நடைமுறைப்படுத்தல்

சொல்வலையை நான்கு ஒழுங்கமைப்புகளாகப் பகுக்கலாம்.

1. சொல்வள ஒழுங்குமுறை (Lexical resource system)
2. ஒருங்கிணைக்கும் ஒழுங்குமுறை (Compiler system)
3. சேகரிப்பு ஒழுங்குமுறை (Storage system)
4. மீளப்பெறும் ஒழுங்குமுறை (Retrieval system)

மீளப்பெறும் ஒழுங்குமுறை பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

1. வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை மின்தட்டச்சு அமைப்பாக மாற்றுகின்றது.
2. வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தைத் தகவல் மீட்பு ஒழுங்கமைப்பாக மாற்றுகின்றது.
3. இணைய மீளப்பெறுதல்: வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை குறி தேடலுடன் இணைப்பது.

10. முடிவுரை

சொல்வலை என்பது சொற்களைக் கருத்துருக்களில் அடக்கி அவற்றை ஒருபொருள் பன்மொழியக் குழுமங்களாகப் (synsets) பகுத்து அவைகளுக்கிடையே உள்ள பொருண்மை மற்றும் சொல் உறவுகளை நிறுவி ஒரு வலையமைப்பாக வடிவமைப்பதாகும். சுருக்கமாகச் சொல்வலை என்பதை விளக்கவேண்டுமானால் அது ஒரு அகராதிப் பண்புக்கூறுகளும் பொருட்புல அகராதிப் (சொற்களஞ்சியம்) பண்புக்கூறுகளும் கொண்ட ஒரு சொல் மூலவளமாகும். கால வளர்ச்சிக்கேற்பச் சொல் விளக்க அகராதிகளும் சொற்களஞ்சியங்களும் கணினிமயமாக்கப்பட்டுச் சொல்வலை என்ற புதிய வடிவைப் பெற்றுள்ளன. சொல்வலை பயன்படுத்துவோர் தேவையான தகவல்களைக் கணினி வழி எளிதாகப் பெறவேண்டும் என்ற நோக்கில் உருவாக்கப்பட்டுள்ளது.

இன்று இணைய தளங்களில் தரப்படும் தரவுகள் நாளுக்கு நாள் விரிவடைந்து வருகிறது. உலகளவு விரிந்த வலை (World Wide Web), இணையதளத் தரவு மையங்கள் மற்றும் நிர்வகிக்கப்பட்ட தகவல் ஒழுங்கமைப்புகள், ஆவணக் காப்பகங்கள் (document archieves) மற்றும் உள்ளிடை இணையங்களின் ((Intranet) வளர்ச்சி என்பன யாவும் தகவல்களின் மூலவளமாகச் சேவை செய்கின்றன. புதிய இடைமுகங்களாலும் கருவிகளாலும் தகவல் அணுகல் மேம்பட்டுள்ளது; ஆனால் தகவலானது இப்போதும் பயன்பாட்டாளரால் தரப்படுகின்ற சில முக்கியச் சொற்களால் (Keywords) இணக்கமானதாக அல்லது பொருத்தமானதாகப் பொதுவாக அடையாளம் காணப்படுகிறது; தகவல் தேடல் சொல்லடைவு செய்யப்பட்ட இந்தச் சொற்கள் மூலமாகவோ இந்தச் சொற்களைக் கொண்ட தலைப்பு அடிப்படையிலோ நடைபெறுகிறது. எதிர்காலத்தில் இம்மூலங்களின் அளவெல்லை மற்றும் தகவலின் அளவு அதிகரிக்கும் போது சரியான தகவலை இடங்காண சரியான கலைச்சொல்லைப் பயன்படுத்துவதற்குத் திறனற்ற பயன்பாட்டாளர் கூடுதல் சிக்கலை எதிர்கொள்வர்.

சொல்வலை இயற்கை மொழி ஆய்வுக்கும் உருவாக்கத்திற்கும் இது பல வழிகளில் கைகொடுக்கும். சொல்வலையைப் பயன்படுத்தி இயந்திரமொழிபெயர்ப்பு செய்வதற்கான முயற்சிகளும் நடைபெற்று வருகின்றன. சொல்வலையில் உட்படுத்தப்பட்டுள்ள சொல் அறிவு அதைத் திறமையான ஒழுங்கு முறைகள் (expert systems), மொழிபெயர்ப்புத் துணைக்கருவிகள், தேடல் இயந்திரங்கள், கற்றல் ஒழுங்கமைப்புகள் மற்றும் தானியங்கு சுருக்கிகள் (automatic summarizers)

போன்றவற்றிற்குப் பொருத்தமானதாகச் செய்கின்றது.

ஐரோப்பாவில் பல மொழிகள் தகவலை இடம் காண்பதில் சிக்கலை நேரிடுகின்றன. யூரோ சொல்வலை சொல் உறவுகளால் தொடர்புப்படுத்தப்பட்ட சொற்களின் பன்மொழி வலையமைப்பைத் தருகிறது. இது தகவல் தேடலில் பயன்படுத்தப்படும் பயன்பாட்டாளரால் தரப்படும் சொற்களிலிருந்து பிற சொற்களைக் கண்டுக்கொள்ளத் தேடல் கருவியைத் தருகிறது. இது குறிப்பாகச் சொற்றொகுதியின் போதிய அறிவு இல்லாத இரண்டாவது மொழியில் பணி செய்யும் பயன்பாட்டாளர்களுக்குப் பயன்படும். இந்தச் சொல்வலை பிற பயன்பாடுகளுக்கு ஆதரவாக இருக்கும் ஒரு அடிப்படை மூலமாகப் பயன்படுத்தப்படுகிறது. சொல்வலையில் உட்படுத்தப்படும் உள்ளுறையும் பொருண்மை அறிவு வலுவான ஒழுங்குமுறைகள், மொழிபெயர்ப்புக் கருவிகள், மொழி கற்கும் ஒழுங்குமுறைகள் மற்றும் தானியங்கி சுருக்கிகள் ஆகியவற்றின் ஒரு பகுதியாகப் பொருத்தமுறச் செய்கிறது.

யூரோ சொல்வலை போலவே திராவிடமொழிகளின் சொல்வலை (Dravidian wordnet) உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்படவேண்டும். தமிழுக்கு என்று தனியான ஒரு சொல்வலை உருவாக்கப்படவேண்டும் என்ற குறிக்கோளை நிறைவு செய்யவேண்டுமானால் அதற்குரிய முன்னோடி ஆய்வுகள் தேவை. சொற்பொருண்மையியல் கோட்பாட்டை முழுவதும் அறிந்து கொண்டு பின்னர் அக்கோட்பாட்டை ஒட்டி பல ஆய்வுகளை மேற்கொண்டு தேவையான மூலவளங்களை உருவாக்கியப் பின்தான் தமிழச் சொல்வலை உருவாக்கம் தொடங்கப்படவேண்டும். அடிப்படையாக ஒரு அகராதியும் பொருட்புல அகராதியும் (சொற்களஞ்சியமும்) தேவை. கிரியாவின் தற்காலத் தமிழகராதியும் (சுப்பிரமணியம், 1992) தற்காலத் தமிழ்ச் சொற்களஞ்சியமும் (இராசேந்திரன், 2001) இதைப் பூர்த்தி செய்கின்றன. மேலும் பல சிறப்பு அகராதிகளும் நிகண்டுகளும் தமிழ் மூலப்பொருண்பற்றிய இராசேந்திரனின் நூல்களும் தரவுமைய உருவாக்கத்திற்குப் பயன்படுத்தப்படவேண்டும்.

தமிழ்ச் சொல்வலையில் அடங்கும் சொற்றொகை சொற்கள் பெயர், வினை, பெயரடை, வினையடை போன்ற சொற்பாகுபாடுகளின் அடிப்படையில் பிரிக்கப்பட்டுப் பயன்படுத்துவோருக்கு எளிமையான முறையில் சொல் குறித்த தேவையான செய்திகளைப் பெறவேண்டி கணினிமயமாக்கப்பட்டு ஒரு இணையதளச் சொல் தரவுதளமாகத் (online lexical database) தரப்படவேண்டும். தமிழ்ச் சொல்வலை பொருண்மை உறவுகளுக்கும் சொற்களுக்கும் இடையிலான உறவுளுக்கும் முக்கியத்துவம் தரவேண்டும். பெயர், வினை, பெயரடை, வினையடை போன்ற வகைபாடுகளில் அடங்கும் சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளை வெளிப்படுத்தும் வகையில் தரவுத்தளம் (database) உருவாக்கப்பட்டு பயன்படுத்துவோர் தேவையான தகவல்களைப் பெறும்படி ஒரு “முன் முகப்பு” உருவாக்கப்பட வேண்டும்.

சொற்பொருண்மையியல் கணிப்பொறியியலுடன் உள்ள தற்போதைய தொடர்புகாரணமாகக் கணினிச் சொற்பொருண்மையியல் என்ற புதிய ஆய்வுக்களமாக மலர்ந்துள்ளது. சொற்பொருண்மையியல் கணினிமொழியியலில் பெரிய ஆய்வுக் களமாக மாறிவருகின்றது. இத்தகைய கணினிமொழியியல் கோட்பாடுகள் அடிப்படையில் அமைவதுதான் சொல்வலை.

தமிழ்ச் சொற்களில் பெயர்ச்சொற்கள், வினைச்சொற்கள், பெயரடைச்சொற்கள், வினையடைச்சொற்கள் ஆகியவை அடங்குகின்றன. பெயரும் வினையும் தலைமை இலக்கணக்கூறுகள் என்றும் பெயரடையும் வினையடையும் துணைமை இலக்கணக்கூறுகள் என்றும் அழைக்கப்படுகின்றன. இச்சொற்களை ஒருபொருள் பன்மொழியக்குழுமங்களாகப் பகுத்து அவற்றிற்கு இடையில் வரும் சொல் மற்றும் பொருண்மை உறவுகள் வெளிப்படும் வண்ணமும் அவற்றின் பல்பொருண்மை வெளிவரும் வண்ணமும் அமைத்துப் பயன்படுத்துவோருக்குச் சொல் குறித்த தேவையான செய்திகளை எளிதாகப் பெற ஒரு மென்பொருளின் மாதிரி உருவாக்கப்படவேண்டும்.

தமிழ்சொற்களைப் பொருண்மைக் களங்களாகப் பகுத்து அவற்றில் வரும் சொற்களைப் பொருண்மை உறவுகளால் தொடர்புபடுத்துவது கடினமான செயல்பாடு. இத்தகைய முயற்சி மேற்கொள்ளப்படவேண்டும். இது ஒரு முழுமையான தமிழ்ச்சொல்வலை உருவாக்கத்திற்கு அடிப்படையாக அமையும். பல்வேறு மொழியியல் அறிஞர்கள் உருவாக்கிய பொருண்மையியல் கோட்பாடு அடிப்படையில் உருவாக்கப்பட்ட இராசேந்திரனின் (இராசேந்திரன், 2001) தற்காலத் தமிழ்ச் சொற்களஞ்சியத்தில் பட்டியலிடப்பட்டுள்ள பெயர், வினை, பெயரடை, வினையடைச் சொற்கள் சொல்வலைக்கு ஏற்பத் தரவுதளமாக மாற்றப்படவேண்டும். இச்சொல்வலை தேவையான மென்பொருள்களைப் பயன்படுத்தி ஒரு மீளப்பெறும் ஒழுங்குமுறையாக உருவாக்கப்படவேண்டும்.

தமிழ்ச் சொல்வலை ஒரு சொல் தரவு மையம் (Lexical database). இதன் முக்கியமான பண்பு அதன் அர்த்தங்களின் வலைப்பின்னல். சொற்களுக்கு இடையே உள்ள ஒற்றுமையை அளந்து சொற்பொருள் மயக்கம் தீர்க்கச் சொல்வலை பயன்படுகின்றது. சொல்வலையைப் பயன்படுத்தி படிநிலை அமைப்பில் தரப்பட்டுள்ள இரு சொற்களுக்கு இடையே உள்ள பொருண்மை உறவை ஆய்ந்து அச்சொற்களுக்கு இடையே உள்ள ஒற்றுமையை அளக்கலாம். பொருண்மை மயக்கம் தவிர மீட்புச்செயல்பாடுகளுக்கும் சொல்வலை பயன்படுகின்றது. தகவல் தளங்களிலும் தகவல் மையங்களிலும் இருந்து பயன்படுத்துபரின் கேள்வியால் தேவையான தகவல்களைப் பெற இயலும். சொல்வலை

இதற்குப் பெரிதும் உதவும். இயற்கை மொழி ஆய்வுக்கும் உருவாக்கத்திற்கும் இது பல வழிகளில் கைகொடுக்கும். சொல்வலையைப் பயன்படுத்தி இயந்திரமொழிபெயர்ப்பு செய்வதற்கான முயற்சிகளும் நடைபெற்று வருகின்றன. சொல்வலையில் உட்படுத்தப்பட்டுள்ள சொல் அறிவு அதைத் திறமையான ஒழுங்குமுறைகள் (expert systems), மொழிபெயாப்புத் துணைக்கருவிகள், தேடல் இயந்திரங்கள், கற்றல் ஒழுங்கமைப்புகள் மற்றும் தானியங்கு சுருக்கிகள் (automatic summarizers) பேன்றவற்றிற்குப் பொருத்தமானதாகச் செய்கின்றது. யூரோ சொல்வலை போலவே திராவிடமொழிகளின் சொல்வலை (Dravidian wordnet) உருவாக்குவதற்கான முயற்சிகள் தொடங்கப்படவேண்டும். தமிழ்ச் சொல்வலையின் மாதிரிதான் உருவாக்கப்பட்டுள்ளது. முழுச்சொல்வலை உருவாக்கப்படவில்லை. தற்போது மனிதவள மேம்பாட்டு அமைச்சகத்தின் நிதி உதவியுடன் தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் நடைபெற்ற திராவிடமொழிகளின் சொல்வலைகள் என்ற திட்டத்தின் ஒரு பகுதியாக அமையும் தமிழ்ச்சொல்வலை உருவாக்கத்திற்கு இது முன்னோடியாக அமையும்.

இந்திய மொழிகளின் தொழில் நுட்பவளர்ச்சி (Technological Development Indian Languages) இணையதளத்தில் இந்தோ சொல்வலையின் ஒருபகுதியான தமிழ் சொல்வலையை பற்றிய விளக்கங்களையும் பயன்படுத்துவர் இடைமுகம் மூலம் சொல்வலைத் தகவல்களையும் பெறலாம்.

துணைநூல்கள்

- இராசேந்திரன் ச. 1991. “தற்காலத் தமிழ்ச் சொற்களஞ்சியம்” (பாகம்1) தமிழ்க் கலை 9.1- 4:55-76.
- இராசேந்திரன் ச 1999. “பொருட்புல வகைப்பாடும் சொற்களஞ்சியமும்”. புலமை 25.2:47-66
- இராசேந்திரன் ச 2001. “தற்காலத் தமிழ்ச்சொற்களஞ்சியம்”. தஞ்சாவூர்: தமிழ்ப் பல்கலைக்கழகம்.
- இராசேந்திரன் ச 2006. தமிழ்ச் சொல்வலை. ஆய்வுமுகங்கள். எச். சித்திரபுத்திரன் மற்றும் பிறர் (பதிப்பு). ஆசிரியர் பேரவை. தமிழ்ப் பல்கலைக்கழகம். தஞ்சாவூர், 2006, 105-117.
- இராசேந்திரன், ச. மற்றும் பாஸ்கரன், ச. 2006. “தமிழ் மின்சொற்களஞ்சியம்”, தஞ்சாவூர்: தமிழ்ப் பல்கலைக்கழகம்.
- இராசேந்திரன், ச. மற்றும் அனிதா, க. 2019. தமிழ்ச் சொற்றொகையின் மூலப்பொருண்மையியல் ஆய்வு (Ontology of Tamil Vocabulary) Language in India www.languageinindia.com ISSN 1930-2940 Vol. 19:8 August 2019.
- இராசேந்திரன், ச. மற்றும் பாஸ்கரன், ச. 2006. “தமிழ் மின்சொற்களஞ்சியம்”, தஞ்சாவூர்: தமிழ்ப் பல்கலைக்கழகம்.
- இராசேந்திரன் ச. 2021 சொல்வலையும் அதன் பரிமாணங்களும் (WordNet and its dimensions). Language in India www.languageinindia.com ISSN 1930-2940 Vol. 21:3 March 2021
- Rajendran. S, 2009. Dravidian WordNet. In: Proceedings of Tamil Internet Conference 2009. Cologne, Germany, October, 2009
- சக்திவேல், சு மற்றும் ச. இராசேந்திரன். 1994. சொற்கள் வாழ்வும் வரலாறும் [Words: Life and History]. மணிவாசகர் பதிப்பகம், சென்னை.
- Miller, G.A. 1991. “Science of Words”. New York: Scientific American Library.
- Miller, G.A.1990. “Nouns in WordNet:a lexical inheritance system”. International Journal of Lexicography Vol 3, No. 4, 245-264.
- Miller G.A. , R. Beckwith, C. Fellbaum, D. Gross, K.J. Miller. 1990. “Introduction to WordNet: An On-line Lexical Database.” International Journal of Lexicography, Vol 3, No.4, 235-244.
- Miller, K.J. 1998. “Modifiers in WordNet”. In: Fellbaum, C. (ed.). 1998. WordNet: “An Electronic Lexical Database”. Cambridge: MIT Press.
- Nair, Nandu C. Rajendran S, Batsure, K. Aligning IndoWordNet with the Princeton WordNet. Down loaded in 2019.
- Rajendran, S. 2002. “Preliminaries to the preparation of Wordnet for Tamil.” Language in India 2:1, March 2002, www.languageinindia.com
- Rajendran, S. 2009. “Dravidian WordNet.” In: Proceedings of Tamil Internet Conference 2009. Cologne, Germany, October, 2009
- Rajendran, S. 2010. “Tamil WordNet.” In: Proceedings of the Global WordNet Conference (GWC 10) 2010, IIT, Bombay.
- Rajendran, S. 2016. “Tamil Thesaurus to WordNet.” In: Conference Papers of 15th Tamil Internet Conference 2016. International Forum for Information Technology in Tamil, September 2016, 1-9.
- Rajendran. S, 2010. Tamil WordNet. In: Proceedings of the Global WordNet Conference (GWC 10) 2010, IIT, Bombay
- Rajendran S. & Anandkumar, M. 2017 Visual Onto-thesaurus for Tamil. Language in India www.languageinindia.com ISSN 1930-2940 Vol. 17:5 May 2017
- Rajendran, S., S. Arulmozi, B. Kumara Shanmugam, S. Baskaran, and S. Thiagarajan. 2002. “Tamil WordNet.” In: Proceedings of the First International Global WordNet Conference. Mysore: CIIL, 271-274.
- Rajendran,S., Shivapratap G, Dhanalakshmi V and Soman K.P. 2010. “Building a WordNet for Dravidian Languages.” In: Proceedings of the Global WordNet Conference (GWC 10), 2010 IIT Bombay
- Vossen, Piek . 1998. Introduction to EuroWordNet. In EuroWordNet: A multilingual database with lexical semantic networks, pages 1-17. Springer.
- Vossen, Piek (eds). 1998. Special Issue on EuroWordNet. Computers and the Humanities, Vol 32, Nos. 2-3, 91-115.
- Vossen P. (eds.) 1999 “EuroWordNet: a multilingual database with lexical semantic networks for European Languages”. Dordrecht: Kluwer Academic Publishers.
- Vossen P. 1999 fc., “EuroWordNet as a multilingual database”. In: Wolfgang Teubert (ed). Berlin: Mouton Gruyter.

NATURAL LANGUAGE PROCESSING FOR TAMIL

Zero-shot Response generation in Chatbots

Sobha Lalitha Devi and Pattabhi RK Rao

ABSTRACT

Response generation is one of the main components in conversational AI. This task involves “understanding" natural language input provided by the user to give an appropriate response in a natural, fluent way as human’s converse, for instance back and forth dialogues. The emergence of large-scale dialogue datasets availability in English has greatly helped in advancing the research of dialog in English and few European languages. In the Indian languages there are no such datasets available. In this work we have created a small Tamil dialogue dataset in general domain. This dataset has 50 conversations between two speakers. In this work a zero-shot algorithm which uses a capsule-based model, as described by Xia (Xia et.al. 2018) is developed for response generation. A F1-accuracy score of 82.13% is obtained which is comparable with the state of the art.

Sobha Lalitha Devi and Pattabhi RK Rao

AU-KBC Research Centre, MIT Campus of Anna University,

Chromepet, Chennai, India.

sobha@au-kbc.org

pattabhi@au-kbc.org

1. INTRODUCTION

Response generation is one of the main components in conversational AI. This task involves “understanding" natural language input provided by the user to give an appropriate response in a natural, fluent way as human’s converse, for instance back and forth dialogues. The creation of high-quality natural language responses for chatbots remains a challenging and time-consuming task that often depends on high-quality training data and deep domain knowledge. Therefore, it is essential to engage experts in the chatbot response development process which have the required domain knowledge. But having a high knowledge human expert available is difficult and expensive. It is essential that automatic or semi-automatic methods for response generation are devised. The success of neural models and the emergence of large-scale dialogue datasets availability in English have greatly helped in advancing the research of dialog generation (Serban et al., 2016, 2017; Huang et al., 2020; Meng et al., 2020) in English and few European languages. In the Indian languages there are no such datasets available. Hence it is important to use methods which can work on limited datasets or practically no annotated datasets.

The paper is further organized as follows: Section 2 describes state of the art in this area of research. Section 3 describes the present work which also includes description of data creation. In section 4, experiments and results are discussed. The paper concludes with section 5.

2. RELATED WORK

Advancements and accessibility of high-end hardware infrastructure has helped researchers use artificial neural network algorithms with success to develop neural models such as BERT, Large Language Models (LLMs). In languages such as English and few European languages emergence of large-scale datasets have greatly advanced the research of conversational AI. The open-domain chatbot systems generate more informative and fluent responses (Ke et al., 2018; Zhang et al., 2020b; Bao et al., 2020; Meng et al., 2021), for general conversations and for domain specific applications such as providing emotional companionship and social chatbots.

It is observed that most of the response generation systems (Zhang et al., 2020b; Bao et al., 2020; Li et al., 2020; Floridi and Chiriatti, 2020) rely on a considerable amount of data resource, such as DialoGPT (Zhang et al., 2020b). But one of the greatest hurdles is that, the dialogue corpus for many languages is unavailable, which limits the usefulness of the available conversational AI systems for low-resource or even zero-resource languages such as Tamil and other Indian languages. There is a need to design and develop approaches that can efficiently perform with limited training corpus. The pre-trained language models such as GPT-3 (Brown et al, 2020), have been used to augment small datasets for the development of chatbots. Even the language models such as GPT-3 have limited usability for our Indian languages as these generate generic responses and sometimes even irrelevant responses. Thus, there is a need to develop Zero-shot Learning for response generation. This task refers to building a response generation system with very few training samples (Floridi and Chiriatti, 2020). Most existing zero-shot methods rely on large-scale pre-trained generative models (Lewis et al., 2020; Zhang et al., 2020b; Floridi and Chiriatti, 2020), such as GPT-3 (Floridi and Chiriatti, 2020). These methods require huge computing resources, which hinders the usableness of response generation systems. Also as stated above use of such LLMs for our Indian languages mostly provide empty responses or irrelevant responses.

3. OUR APPROACH

In this work we have created a small dialogue dataset in general domain. This dataset has 50 conversations between two speakers. And develop a base response generation system using neural methods. Since the data is very small, in the real time scenario of chatbot we need to handle unseen user inputs for responses need to be generated by the system. Here we tackle this issue using zero-shot algorithm which uses a capsule based (Hiton et.al. 2011, Sabour et.al. 2017) model, as tried by (Xia et.al. 2018).

3.1 Data Creation

We have developed conversation (or dialogue) dataset in-house. In developing the corpus, we have followed the annotation convention used for developing annotated corpus of free conversations in Japanese, called "JAIST Annotated Corpus of Free Conversations" (Kiyoaki Shirai and Tomotaka Fukuoka. 2018). Our corpus consists of dialogs of two native speakers in Tamil as participants, where they freely talk about various topics. Each utterance in the dialogs is annotated with two kinds of tags. One is a Turn construction Unit (or speech act), which is the type of utterance that represents the speaker's intention. The other is sympathy that is the interest shown by the listener in the current topic in the conversation (response). The corpus consists of transcriptions of 50 free conversations between two participants. The total duration of the dialog is about 50 hours. Each utterance was transcribed by hand. Out of 50 conversations, not all were with two participants. 46 dialogs were with two people participate in the conversation. The statistics is given: Number of dialogs 46, Number of utterances 23556, Average number of utterances per dialog 500. This shows that each dialog is long. We had two annotators and the Inter annotator's agreement Kappa score is 92%. This Tamil conversational data is first of its kind in Indian languages.

Each utterance has the information: Speaker ID: An identification number of the speaker. Turn taking: A flag indicating whether the speaker has changed or not. TCU: A dialog act of an utterance. Sympathy tag: A tag that represents whether the listener shows sympathy or antipathy. Nine types of TCU were formulated for the annotation (Request, Confirmation etc.). The annotation has also three tags for sympathy.

3.2 CapsNet - System Implementation

In this work two tier architecture approach similar to the one proposed by Xia et al, (2018) is followed. Initially the CapsNet model is developed using our small dataset to generate responses. Then the Zero-shot learning for generation of responses for unseen responses is used.

The steps in the implementation of CapsNet response generation are described below:

a) Dataset preparation:

- We have developed labelled dataset for response generation. Each utterance is labelled with fundamental utterance unit the Turn Construction Units (TCUs) and the complete utterance is labelled with its corresponding class of the response.
- Pre-process the conversation data by performing proper tokenization and other typological error corrections, as this is a transcribed data.

b) CapsNet Data Input

- Convert the pre-processed conversation data into vector representations such that it can be fed into the neural network. Word embedding's method of Word2Vec is used to represent TCUs as dense vectors.

c) Designing the CapsNet architecture for response generation:

- The original CapsNet architecture used for image recognition as used by Hinton

et al 2017 is adapted to suit the response generation task. The image-based input is replaced with the text vector representations obtained from Word2vec.

- The number of capsule layers is derived empirically.
- Dynamic routing-by-agreement algorithm (Sabour et al., 2017), is used for

$$q_{lk} = \frac{exp\{-d(\mathbf{e}_{z_l}, \mathbf{e}_{y_k})\}}{\sum_{k=1}^{K} exp\{-d(\mathbf{e}_{z_l}, \mathbf{e}_{y_k})\}},$$

where

$$d(\mathbf{e}_{z_l}, \mathbf{e}_{y_k}) = (\mathbf{e}_{z_l} - \mathbf{e}_{y_k})^T \Sigma^{-1} (\mathbf{e}_{z_l} - \mathbf{e}_{y_k})$$

Learning hierarchical relationships between TCUs and thus used for response generation.

d) Training the CapsNet:

- The dataset is split into training and validation sets.
- CapsNet model is initialized with initial values and then use categorical cross-entropy loss function.
- Train the model using the training dataset, monitoring the validation accuracy to avoid over-fitting.
- Backpropagation and gradient descent function to update the model's weights and thus optimize the loss function.

e) Measuring response relations:

- Here we propose to learn a Mahalanobis distance metric to measure the relationship between unseen and seen responses. Specifically, given the embeddings of an unseen response l and a seen response k, their squared Mahalanobis distance is given by:

  dM (eul , esk) =

  (eul − esk)>Ω−1(eul − esk), ---- (1)

- Where, Ω is a learnable covariance matrix which models the correlation between dimensions of the embedding. Note that CapsNet (Xia et al., 2018) also tries to use Eq. (1) to model the relationship between unseen and seen responses, but it ignores the correlation between dimensions and simply sets Ω = σ2I (σ is a scaling hyper-parameter), which is a scaled squared Euclidean distance.

3.3 Zero-shot Response Generation

Zero-shot response generation involves using auxiliary information or semantic descriptions to associate visual features with labels. The zero-shot learning utilizes vote vectors from existing responses to build response representations for emerging new responses via a similarity metric between unseen intents and responses.

Suppose there are K existing (seen) responses and L emerging (unseen) responses, the similarities between existing and emerging responses form a matrix Q∈R L×K. Specifically, the similarity between an emerging response zl ∈Z and an existing yk ∈Y is computed as ezl , eyk ∈ RD 1×1 are response embeddings computed by the sum of word embeddings of the response label. Σ models the correlations among response embedding dimensions and we use Σ = σ2 I. σ is a hyper-parameter for scaling.

We feed the prediction vector nl to Dynamic routing algorithm and derive activation vectors nl on emerging responses as the output. The final response representation nl for each emerging response is updated toward the direction where it coincides with representative votes vectors. We can easily classify the emerging responses by choosing the activation vector with the largest norm $\hat{Z}$ = arg max ||n1 ||

4. EXPERIMENTS AND RESULTS

The experiments are performed using the data created by us which was described in the section 3.1. The data is split into two, training (38 dialogues) and test (8 dialogues). The test partition is formed such that 4 dialogues have topic similarity with the training partition. The remaining 4 dialogues are completely different than the rest (completely unseen).

The embeddings needed for the response generation models, are developed using Tamil Wikipedia content and copyright free Novels digitized content from Project Madurai (Project Madurai). These pre-trained word embeddings are used in the training of CapsNet as well as for Zero-shot learning as augmentation to the dataset.

A three-fold cross-validation to choose hyper parameters is performed. The dimension of the prediction vector DP is 10. DI = DW because we use the averaged word embeddings contained in the intent label as the intent embedding. An additional input dropout layer with a dropout keep rate 0.8 is applied to the intent annotated corpus of ours. In the loss function, the down-weighting coefficient − λ is 0.5, margins m+ k and mk are set to 0.9 and 0.1 for all the existing intents.

The iteration number iter used in the dynamic routing algorithm is 3. Adam optimizer is used to minimize the loss function.

We have evaluated our method using different models, we employ both automatic metrics and human evaluations.

Automatic Metrics: We employ perplexity (PPL) and distinct 1/2 (Dist.1/2) following previous studies (Zhang et al., 2018; Zheng et al., 2020; Song et al., 2021). Lower perplexity means more reliable model. Distinct 1/2 (Li et al., 2016) are the ratio of distinct uni-grams / bi-grams. Higher distinct means better diversity of responses generated by the model.

Method	PPL	Dist.1/2
TF-IDF Classifier (Base system)	140.65	11.59/43.68
Zero-shot Response generator	110.65	13.59/43.68

Human Evaluation: We further conduct human evaluations to assess the proposed learning framework. We ask three graduate students to evaluate the quality of generated responses for 100 randomly sampled input contexts. We request evaluators to choose a preferred response in a scale of 1 to 5, considering the following aspects of response quality: fluency, informativeness, coherence, and engagingness. We have obtained encouraging results. We have obtained human evaluation score of 82.13%.

Limitations:

CapsNet (Xia et al., 2018) is the first work to employ capsule networks for zero-shot learning. It exploits the self-attention mechanism to extract semantic features (capsules) of an utterance. For zero-shot intent classification, it utilizes the vote vectors of seen responses and the similarities between seen and unseen responses based on Euclidean distance to make predictions for unseen new responses. CapsNet has demonstrated strong performance, but has two fundamental limitations:

i. The self-attention module of CapsNet has little issues in handling the polysemy problem, which limits the representation capacity of semantic capsules. This needs more cautious implementation.

ii. For the generalized zero-shot classification setting, the method of CapsNet for constructing the prediction vectors is highly likely to cause the model to lose generalization ability to unseen intents

Ethics Statement

We have ensured that all the volunteers from whom the conversation data is collected have been well informed the purpose of the work was clearly explained and prior consent of the speakers was obtained and the data is anonymized by removing all sensitive personal information such as speaker's names.

Acknowledgments

We thank all the volunteers for giving their consent and participation in the data collection work.

REFERENCES

- Anjishnu Kumar, Pavankumar Reddy Muddireddy, Markus Dreyer, and Bj¨orn Hoffmeister. 2017. Zero shot learning across heterogeneous overlapping domains. In Annual Conference of the International Speech Communication Association (INTER-SPEECH), pages 2914–2918.
- Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In Annual Conference of the International Speech Communication Association (INTER SPEECH), pages 685–689.
- Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S. Yu. 2018. Zero-shot user intent detection via capsule neural networks. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3090–3099
- Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lef`evre. 2015a. Online adaptative zero-shot learning spoken language understanding using wordembedding. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5321–5325.
- Emmanuel Ferreira, Bassam Jabaian, and Fabrice Lef`evre. 2015b. Zero-shot semantic parser for spoken language understanding. In Annual Conference of the International Speech Communication Association (INTERSPEECH), pages 1403–1407.
- Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In ICANN, pages 44–51.
- Jian Hu, Gang Wang, Frederick H. Lochovsky, JianTao Sun, and Zheng Chen. 2009. Understanding user's query intent with wikipedia.

In International Conference on World Wide Web (WWW), pages 471–480.

- Jinseok Nam, Eneldo Loza Menc'ia, and Johannes F¨urnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In AAAI Conference on Artificial Intelligence (AAAI), pages 1948–1954.
- Kiyoaki Shirai and Tomotaka Fukuoka. 2018. JAIST Annotated Corpus of Free Conversation. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).
- Prateek Veeranna Sappadla, Jinseok Nam, Eneldo Loza Menc'ia, and Johannes F¨urnkranz. 2016. Using semantic similarity for multi-label zero-shot classification of text documents. In European Symposium on Artificial Neural Networks (ESANN).
- Suman Ravuri and Andreas Stolcke. 2015. Recurrent neural network and lstm models for lexical utterance classification. In Annual Conference of the International Speech Communication Association (INTER-SPEECH), pages 135–139.
- Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In Advances in Neural Information Processing Systems (NIPS), pages 3859–3869.
- Sepp Hochreiter and J¨urgen Schmidhuber. 1997. Long short-term memory. Neural computation, 9(8):1735–1780.
- Puyang Xu and Ruhi Sarikaya. 2013. Convolutional neural network based triangular CRF for joint intent detection and slot filling. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU Workshop), pages 78–83.
- Yun-Nung Chen, Dilek Z. Hakkani-T¨ur, and Xiaodong He. 2016. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6045–6049.

Bridging the Linguistic Gap: A Practical Exploration of Tamil Integration in Technology - A Data-Driven Perspective

Somasundaram Meenakshisundaram

ABSTRACT

The digital revolution presents both opportunities and challenges for Tamil still being a low resource language. This research paper examines the practical approaches to bridge the linguistic gap between Tamil and technology through data-driven analysis. It explores the ongoing debate regarding Tamil keyboard standardization, the rise of Tanglish (Tamil-English hybrid) usage, and the potential of voice-based technologies in fostering Tamil inclusion. Through analysis of successful case studies like Hinglish integration and Singapore's GovTech initiatives, the paper proposes a multi-pronged strategy for enhancing Tamil accessibility in the digital domain.

Somasundaram Meenakshisundaram
CEO, Nayamsoft India Private limited
Chair, Tamil Technology Development Committee, SICCI.
Email: somz@nayamsoft.com

INTRODUCTION

With over 80 million speakers worldwide, Tamil is still facing the complexities of the digital age. Integrating Tamil into technology holds immense potential for cultural preservation, educational empowerment, and economic growth. However, bridging the linguistic gap requires thoughtful consideration of practical challenges and innovative solutions.

Tamil Keyboard Standardization:

Tamil Keyboard Standardization is pivotal for a seamless digital experience in the Tamil language. While initiatives like the Ka-naada keyboard have made progress, the urgent need for a standardized Tamil keyboard raises questions about cultural preservation and technological advancement.



Source: https://yourstory.com/2018/12/indic-keyboard-for-indian-languages-kanaada-guru-prasad

Current Scenario:

Current research emphasizes the significance of standardized Tamil keyboards, with solutions like Ka-naada keyboard which supports Indic language typing showcasing potential. However, the lack of a unified standard raises concerns about future generations losing touch with the traditional Tamil script typing.

Own Tamil Keyboard vs. Existing Solutions:

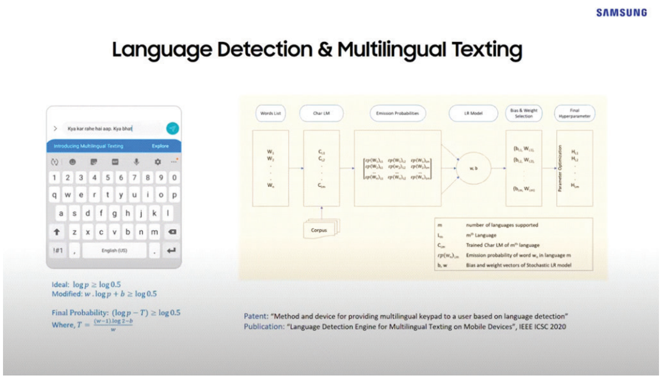
The decision between creating a new Tamil keyboard and adopting existing solutions involves weighing technological implications, user adoption rates, and cultural considerations. Striking a balance between innovation and cultural preservation is imperative for successful standardization.

Samsung R&D's Contribution

Samsung R&D's extensive efforts in Indian language technology play a crucial role in shaping Tamil integration. Advancements in keyboard technology, language models, and AI showcase the potential for industry leaders to drive regional language integration.

Samsung's Commitment to Indian Languages:

Samsung's commitment to supporting over 40 Indian languages demonstrates industry leadership. Their advancements in technology create a roadmap for other tech giants, emphasizing the role of corporate innovation in regional language integration.



Source: https://www.youtube.com/watch?v=r6XOCSD6K6M

(Indian Language Input on Mobile - Challenges and Advances, Mr. Barath Raj, Samsung R&D Institute)

Indian users are multilingual, while we focus on developing pure language solutions, for the mainstream use cases the multi-lingual approach needs to be taken for the users to start using the solution. This might sound like a controversial statement but from the industry point of view any solution that was not developed for practical use cases will have very little adaption.

The landscape of Tamil keyboards reflects a fragmented ecosystem, with solutions like Ka-naada and Tamil 99 gaining traction but failing to achieve universal adoption. The lingering absence of a standardized layout necessitates a critical juncture – should we strive for a unified keyboard for future generations, or remain tethered to existing options? This question holds paramount importance, as the chosen path will impact both the ease of learning and the digital preservation of the Tamil language.

Tanglish: Challenge or Opportunity?

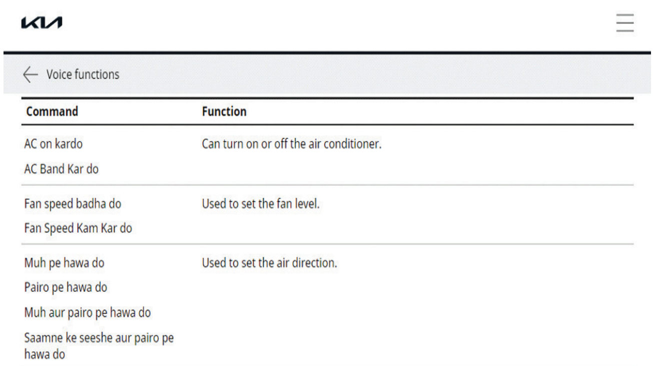
The ubiquitous presence of the QWERTY keyboard has inadvertently birthed a hybrid language phenomenon – Tanglish, a blend of Tamil and English widely used for its ease and familiarity. To ignore this reality would be to marginalize a significant portion of Tamil speakers in the digital space. Therefore, solutions must be sought that acknowledge the prevalence of Tanglish while paving the way for proper Tamil usage. Intelligent autocorrect tools that recognize and translate Tanglish into grammatically accurate Tamil sentences can bridge this gap, fostering a gradual transition towards embracing the richness of the native language.

Hinglish: A Blueprint for Tanglish Integration in Tamil Technology

The digital revolution has not only reshaped communication but also challenged notions of linguistic purity. Across India, the rise of Hinglish – a dynamic blend of Hindi and English – holds valuable lessons for integrating Tanglish, the Tamil-English hybrid, into technology. Analyzing the success of Hinglish reveals a crucial truth: acknowledging user language preferences and embracing hybrid forms can bridge the convenience gap and empower communities.

Hinglish: Paving the Way for Tanglish

Hinglish, no longer relegated to casual conversations, has become a powerful force in Indian technology. Platforms like HaptikAI's virtual assistants understand and respond to Hinglish queries, simplifying customer interactions. Google Pay's Hinglish interface demystifies financial transactions, empowering those less comfortable with pure English. Even carmakers like Kia and MG Motor India have incorporated Hinglish voice commands in their in-car systems, recognizing the growing comfort level of users with this hybrid language. The data behind these successes resonates loud and clear – acknowledging user language preferences through features like context-aware translation, intelligent language recognition, and hybrid interfaces leads to significantly improved user engagement and adoption.



Command	Function
AC on kardo AC Band Kar do	Can turn on or off the air conditioner.
Fan speed badha do Fan Speed Kam Kar do	Used to set the fan level.
Muh pe hawa do Pairo pe hawa do Muh aur pairo pe hawa do Saamne ke seeshe aur pairo pe hawa do	Used to set the air direction.

Source: http://webmanual.kia.com/STD_GEN5W/AVNT/IND/English/008_VR_voicerec.html#d2e14466

Adapting the Hinglish Playbook for Tanglish:

The lessons learned from Hinglish integration offer a clear roadmap for Tanglish. Imagine a Tamil chatbot that seamlessly understands and responds to queries in both languages, bridging the gap for those new to digital platforms. Picture a Google Pay interface in Tamil and English, making financial management accessible to a wider audience. Consider e-commerce platforms like Flipkart incorporating Tanglish voice search, opening up the world of online shopping to users comfortable with this hybrid form. The possibilities are endless, and the potential impact is transformative.

Beyond Convenience: Empowering Communities

Integrating Tanglish is not just about convenience; it's about empowering communities. By acknowledging this hybrid language, we validate the linguistic choices of millions of Tamil speakers and create a digital space that reflects their lived reality. This fosters inclusivity and confidence, encouraging hesitant users to engage with technology in a language they understand and feel comfortable with. Moreover, it preserves and celebrates the unique linguistic tapestry of Tamil culture, ensuring its survival and evolution in the digital age.

CHALLENGES AND CONSIDERATIONS:

Embracing Tanglish is not without its challenges. Concerns regarding standardization, potential dialectal variations, and ensuring accurate interpretation cannot be ignored. However, with careful consideration and collaboration between linguists, technologists, and users, these challenges can be overcome. Leveraging the advancements in Natural Language Processing and machine learning can pave the way for robust Tanglish recognition and translation systems. Open dialogues with the Tamil-speaking community will ensure that these solutions are culturally relevant and user-centric.

EMBRACING THE HYBRID FUTURE

The digital future of Tamil lies not in clinging to linguistic purity but in embracing the vibrant reality of Tanglish. By learning from the success of Hinglish and adopting a data-driven, user-centric approach, we can build technology solutions that empower Tamil speakers, bridge the convenience gap, and unlock a truly inclusive digital experience. Let us weave a digital tapestry that reflects the richness and diversity of the Tamil language, ensuring its continued evolution and prosperity in the ever-evolving landscape of technology.

Tanglish Dominance: A significant percentage of digital users prefer Tanglish over Tamil due to keyboard confusion and familiarity with English keyboards. Just as Hinglish dominates in Hindi-speaking regions, Tanglish is the go-to for many Tamil speakers in digital interactions. By embracing Tanglish, we don't diminish the importance of formal Tamil; we simply acknowledge the reality of our digital landscape. Studies have shown that Tanglish is prevalent in online chats, social media interactions, and even informal writing among Tamil speakers. Ignoring this reality would be akin to building a bridge that only reaches half the population. We build a bridge, not a wall.

Why Multimodality Matters: A User-Centric Approach

Developing technology solutions solely in Tamil, while commendable for language preservation, misses out on the vast section of users who navigate between both languages seamlessly. A multi-modal approach acknowledges this reality and builds applications that cater to both Tamil and English, often within the same user interface. This could involve:

- Hybrid interfaces: Menus, prompts, and instructions offered in both languages, allowing users to switch between them based on their preference.
- Intelligent language recognition: Systems that understand and respond to both Tamil and English inputs, seamlessly interpreting Tanglish phrases and providing accurate response in the chosen language.
- Context-aware translation: Tools that automatically translate specific words or phrases within a sentence, offering users the flexibility to express themselves in their preferred blend of languages without hindering communication.

The Business Case for Multimodality: Beyond Language Purity

From an industry perspective, embracing multimodality is not just about inclusivity, it's a sound business decision. By catering to the linguistic preferences of a broader user base, technology developers can significantly increase their reach and user engagement. Studies have shown that users are more likely to adopt and interact with applications that provide seamless language switching and Tanglish support. This translates to increased user satisfaction, higher retention rates, and ultimately, greater market share.

VOICE: THE INCLUSIVE FUTURE?

How Voice Technology Bridges the Linguistic Gap

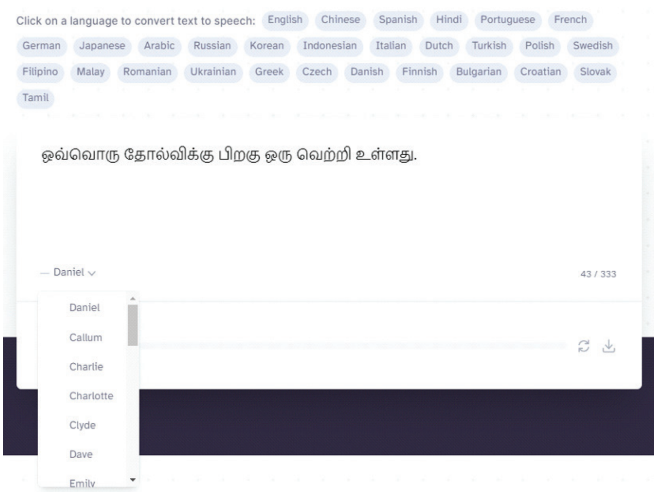
The digital revolution has transformed communication, but for many, linguistic barriers still stand in the way. Enter the realm of voice technology, where the spoken word transcends the limitations of keyboards and screens, offering a beacon of inclusivity for diverse languages like Tamil. This section delves into the immense potential of voice AI for bridging the linguistic gap and enriching the digital experience for Tamil speakers.

Inclusive Communication through Voice Interfaces:

Imagine a world where interacting with technology requires no keyboard, no deciphering complex interfaces, just the natural flow of spoken language. This is the promise of voice technology, and for languages like Tamil, it holds revolutionary potential. Solutions like Eleven Labs and Slang Labs showcase the maturity of voice AI for Tamil, offering near-flawless text-to-speech synthesis, speech-to-speech recognition, and even voice cloning capabilities. This eliminates the need for literacy, opens doors for non-traditional users, and creates a truly inclusive digital platform accessible to all.

Elevating the Auditory Experience:

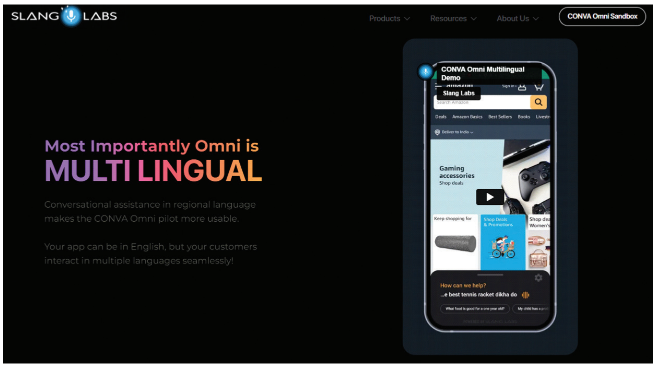
The power of voice technology goes beyond mere accessibility. Platforms like Eleven Labs offer voice cloning features, allowing users to create custom-tailored digital assistants that speak in their own voice or the voice of a beloved public figure. Imagine textbooks narrated in the melodious tones of a favorite writer, news updates delivered in the familiar cadence of a trusted community leader, or virtual assistants responding to queries in the comforting voice of a family member. This personalized auditory experience fosters deeper engagement, strengthens cultural connection, and elevates the digital interaction to a whole new level.



Source: https://elevenlabs.io/text-to-speech

Empowering Communities with Voice Solutions:

Beyond individual experiences, voice technology holds immense potential to empower entire communities. Slang Labs' Conva Omni-Assistant platform transcends pre-built apps, allowing developers to create custom voice solutions for specific needs. Imagine voice-enabled agricultural extension services guiding farmers in their native language, voice-controlled healthcare information systems facilitating medical access in rural areas, or voice-driven educational tools breaking down literacy barriers for underprivileged communities. These are just a glimpse of the possibilities, where voice technology becomes a bridge to knowledge, healthcare, and progress, empowering Tamil communities on a larger scale.



Source: https://www.slanglabs.in/conva-omni-assistant

CHALLENGES AND THE ROAD AHEAD:

While the promise of voice technology for Tamil is undeniable, challenges remain. Dialectal variations within the language must be accounted for, ensuring accurate speech recognition and synthesis across diverse regions. Building robust voice datasets that represent the richness and complexity of Tamil culture is crucial for developing culturally relevant and effective solutions. Collaboration between linguists, technologists, and user

communities will be key to addressing these challenges and shaping voice technology that truly caters to the needs of Tamil speakers.

A FUTURE RESONANT WITH TAMIL VOICES

The digital future of Tamil lies not just in preserving its written form but also in amplifying its vibrant oral traditions through voice technology. By embracing advancements in voice AI, we can dismantle linguistic barriers, empower communities, and create a truly inclusive digital landscape where the diverse voices of Tamil resonate loud and clear. Let us pave the way for a future where technology speaks the language of the people, fostering cultural connection, enriching experiences, and ultimately, celebrating the beauty and power of spoken Tamil in the digital age.

GovTech Initiatives in Singapore

Learning from the Lion City: Singapore's GovTech Blueprint for Tamil-Centric Solutions

Singapore, the gleaming jewel of Southeast Asia, stands as a beacon for successful multilingual GovTech integration. For Tamil speakers in India, its initiatives offer a treasure trove of lessons that can be applied to create a more inclusive and accessible digital experience for Tamil Nadu citizens. Let's delve deeper into Singapore's GovTech landscape and explore its potential to guide the development of Tamil-centric solutions in India.

Singlish Voice Bots: Bridging the Communication Gap

Imagine Siri or Alexa seamlessly understanding and responding to your queries in Singlish, the vibrant hybrid of English and Singaporean Malay. This is the reality in Singapore, where voice bots powered by advanced natural language processing (NLP) have revolutionized citizen engagement. For Tamil Nadu, similar solutions can be developed, empowering individuals comfortable with Tanglish to interact with government services in a familiar and convenient manner. Imagine a healthcare chatbot guiding patients through appointments in both Tamil and English, or an educational assistant offering personalized learning support in a blend of languages. The possibilities are endless, and the impact profound.

Unified DPG Forms Framework: Streamlining Service Delivery

Government forms can be a daunting labyrinth, often riddled with complex jargon and inaccessible language. Singapore's Unified DPG forms framework tackles this challenge head-on by offering a standardized format and supporting Tamil alongside English. This simplifies the application process for citizens, improves transparency, and fosters trust in government services. Adapting this framework to Tamil Nadu's context would be a significant step towards digital inclusivity, ensuring that all citizens have equal access to government services regardless of their language proficiency.



Source: https://vouchers.cdc.gov.sg/residents/how-to-claim-cdc-vouchers-tamil/

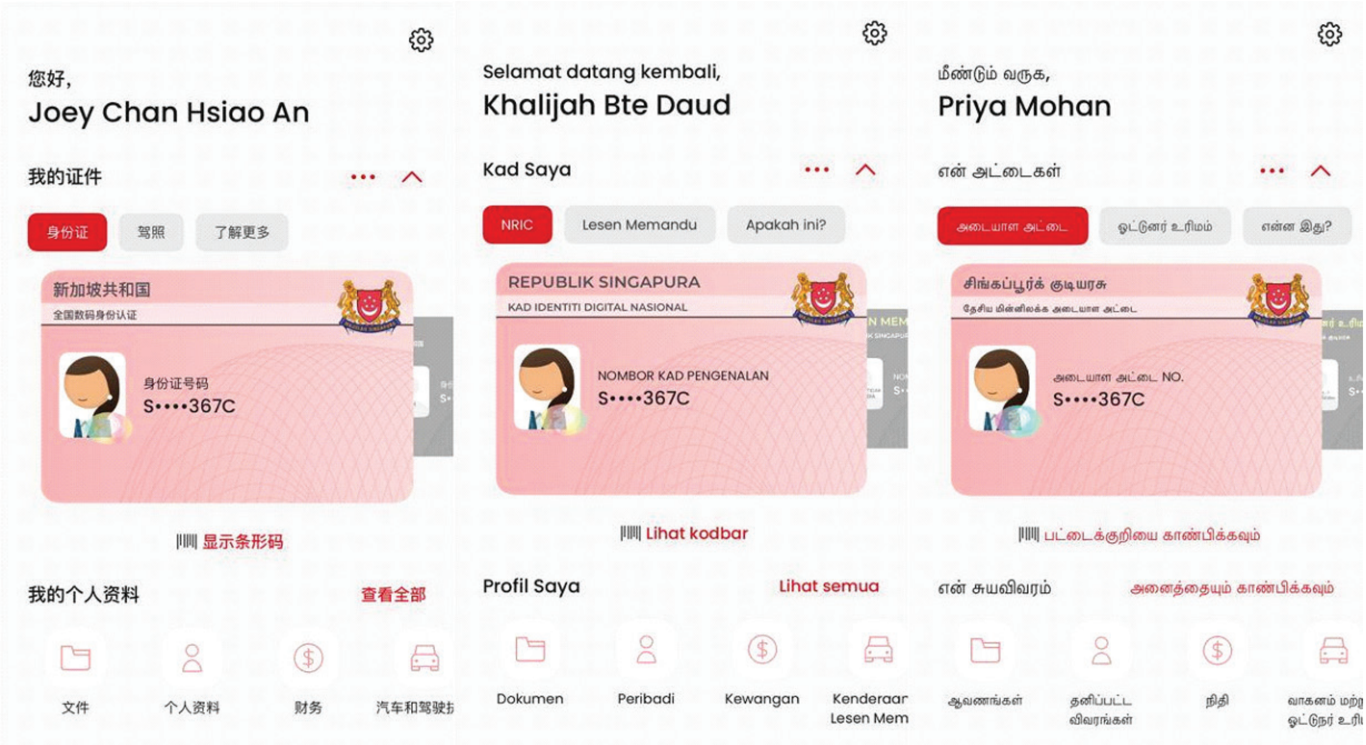
Low-Code Platform: Empowering Citizen Developers

Singapore's low-code platform empowers government agencies and even citizen developers to build digital solutions without extensive coding knowledge. This platform, crucially, supports Tamil, allowing Tamil-speaking developers to create applications and services tailored to their community's needs. Imagine village councils using the platform to develop apps for local resource management, or farmers accessing crop information and market updates in their native tongue. The potential for citizen-driven innovation is immense, and fostering a Tamil-enabled low-code platform can unleash the creativity and problem-solving skills of the Tamil-speaking community.

Data-Driven Insights: From Feedback to Improvement

Singapore's GovTech initiatives are not static; they evolve and adapt based on user feedback and data analysis. The Citizen Translators program, a dedicated group of native speakers who review and refine translations, is a testament to this commitment to continuous improvement. Similarly, Tamil Nadu can

leverage its existing GovTech data, such as bilingual government orders (G.O.), to develop a small language model. This model could be used to personalize interactions, improve the usability of applications, and ultimately deliver a more satisfying digital experience for Tamil speakers.



Source: https://www.tech.gov.sg/media/technews/how-singpass-learnt-three-languages

THE SINGAPORE MODEL: A PATHWAY TO INCLUSIVITY

Singapore's GovTech success story offers a roadmap for Tamil Nadu to follow. By embracing multilingual solutions, streamlining service delivery, empowering citizen developers, and prioritizing user feedback, Tamil Nadu can create a digital ecosystem that truly caters to all its citizens. This is not just about technology; it's about inclusivity, empowerment, and ensuring that everyone, regardless of their language background, has the opportunity to participate in the digital revolution.

CONCLUSION:

Bridging the linguistic gap between Tamil and technology requires a collaborative and multifaceted approach. While the lack of standardized keyboards presents a challenge, exploring potential solutions, including the adaptation of existing layouts or the development of a new one informed by user preferences, holds immense promise. Recognizing the prevalence and convenience of Tanglish necessitates the development of intelligent tools that support this hybrid form while gradually leading users towards proper Tamil usage. Drawing inspiration from the success of Hinglish and the inclusivity of voice technology further illuminates the path forward. Finally, learning from successful GovTech initiatives like those in Singapore can guide the development of Tamil-centric solutions that empower and engage communities. By embracing innovation, collaboration, and a data-driven approach, we can create a future where the vibrant tapestry of the Tamil language seamlessly enriches the digital landscape, ensuring inclusivity and cultural preservation for generations to come.

Bridging Language Barriers for Tamil Travelers Exploring Diverse Regions of India using Deep Learning Technologies

Ra.K.Saravanaguru, Chellatamilan T, Kumar K, Sathyarajasekaran K,

ABSTRACT

Travel is a transformative experience, yet linguistic diversity often poses a challenge for Tamil travelers exploring different regions of India. This research addresses the need for effective communication mechanisms, leveraging deep learning and generative AI (GenAI) technologies to enhance language understanding and facilitate seamless interactions. This study focuses on the development of innovative solutions tailored to the linguistic landscape of India. First, we employ neural machine translation (NMT) models, specifically fine-tuned for languages prevalent in regions frequently visited by Tamil travelers. These models, built on Transformer architectures, ensure accurate and context-aware translations. Complementing this, we introduce multilingual chatting powered by natural language processing (NLP) and sequence-to-sequence models, providing real-time assistance in both Tamil and other Indian languages. To address spoken communication, we integrate speech-to-text and text-to-speech systems, allowing travelers to engage in spoken conversations with locals. Context-aware language models, based on advanced frameworks like Generative Pre-trained Transformer (GPT), enrich translations by capturing cultural nuances and idiomatic expressions. Augmented Reality (AR) translation adds a visual layer to linguistic understanding, overlaying translated text on real-world objects through Smartphone cameras. Interactive language learning apps, incorporating gamified GenAI modules, empower travelers with basic phrases and expressions, fostering a proactive approach to linguistic challenges. Additionally, community-sourced translation platforms engage users in contributing and validating translations, creating a collaborative environment that continuously improves accuracy. These mechanisms not only break down language barriers but also contribute to a more immersive and enriching travel experience. As the research evolves, user feedback and continuous refinement will further enhance the adaptability and effectiveness of these language bridging solutions, ensuring that Tamil travelers can explore any place of India with confidence using deep learning technologies.

Ra.K.Saravanaguru, Vellore Institute of Technology, Vellore.

Chellatamilan T, Vellore Institute of Technology, Vellore,

Kumar K, Vellore Institute of Technology, Vellore.

Sathyarajasekaran K, Vellore Institute of Technology, Chennai.

Corresponding Author: kkumar@vit.ac.in

1. INTRODUCTION

According to the Ministry of Tourism of India, the country's revenue from international tourism increased from 8.7 billion U.S. dollars in 2021 to 16.92 billion dollars in 2022. In 2021, Tamil Nadu had the most domestic tourists among all the states, with more than 115 million visits. The country had a total of over 677 million visits from domestic tourists that year. This financial trajectory mirrors the profound transformation within the tourism landscape. For travelers, the integration of virtual assistants, readily available AI-driven travel agents, and a comprehensive overhaul of the user experience signifies a paradigm shift. From redefining the translation process to enhancing on-the-go engagement, AI based applications are reshaping the very essence of how individuals plan and experience their journeys.

India, with its linguistic diversity and different languages and cultures in different regions, poses many challenges. However, travelers can overcome these challenges using translation mobile apps. These apps use conversational AI chatbots that can converse with locals in any language. The paper presents a voice/text-based conversational AI via chatting, an AR layer to provide linguistic insights, and introduces a gamified GenAI that leverages user feedback to generate community-sourced translations. This system enables Tamil speakers to interact with speakers of other languages without a shared language during their travel to any place in India and reset of the world.

2. BACKGROUND STUDY

This neural machine translation (NMT) model with transformers is a type of natural language processing (NLP) that uses deep neural networks to translate text from one language to another. NMT has achieved remarkable results in recent years, based on the advances in neural network architectures, such as transformers, and large-scale parallel corpora.

However, NMT still faces some challenges, such as handling low-resource languages, preserving discourse coherence, adapting to different domains and styles, and generating diverse and natural outputs. To address these challenges, this work has proposed various models that leverage additional information.

For travelers, the integration of virtual assistants, readily available AI-driven travel agents, and a comprehensive overhaul of the user experience signifies a paradigm shift. From redefining the translation process to enhancing on-the-go engagement, AI based applications are reshaping the very essence of how individuals plan and experience their journeys. This statement is supported by various research papers that explore the impact and potential of AI in the travel industry. For instance, Sia et al. (2023) provide a systematic review of mobile travel applications and their smart features and challenges, highlighting the role of AI in personalizing travel planning and online customer service. Bulchand-Gidumal (2022) discusses how AI can improve travel, tourism, and hospitality by offering relevant offers, reducing costs, and generating more revenue. Li et al. (2019) examines the technology and economic determinants of crypto currency exchange rates, which can affect the payment methods and preferences of travelers. Kirilenko et al. (2018) compare different approaches to automated sentiment analysis in tourism, which can help understand customer reviews and social media posts. These papers demonstrate the diverse and innovative applications of AI in travel, as well as the challenges and opportunities that lie ahead. Additionally, we will explore the background study that is the focal point of this work.

2.1 Multilingual Chatbot

Arivazhagan et al. (2019) introduced a universal Neural Machine Translation (NMT) system capable of translating between any languages pair, handling 103 languages trained on over 25 billion examples. Aharoni et al. (2019) presented an in-depth analysis of existing literature on Multilingual Neural Machine Translation (MNMT), categorizing various approaches based on their central use-case. Kumar and Kumar (2018) explored the avenues of teaching computers to process natural language text by developing a chatbot, appreciating the processes, techniques, the power, and possibilities of natural language processing using recurrent neural networks (RNN). These works provide a comprehensive understanding of the use of NMT in building multilingual chatbots.

2.2 Context Aware Language Models

Context aware language models are a subtopic that focuses on using NMT to generate text that is relevant and coherent with the given context. Context can include various factors, such as the topic, the tone, the style, the user profile, and the history of the conversation. Context-aware language models using Neural Machine Translation (NMT) have been studied extensively. Wu et al. (2022) used BERT to encode contextual information for NMT in a study. The best translation results were obtained by encoding all contextual sequences as one long sequence with BERT. Sugiyama (2021) proposed a simple yet effective NMT approach to context-aware using two primitive components, a sentence-level NMT model and a document-level language model (LM). Another study suggested that context-aware NMT models working on a concatenation of consecutive sentences perform better, but are computationally expensive. These studies highlight the importance and challenges of context-aware language models in NMT.

2.3 Augmented Reality (AR) Translation

The paper "Tourist Experiences through Mobile Augmented Reality" discusses the potential and the correct approach for the implementation of AR in the tourism sector. Another study, "Innovations in Tourism Industry & Development Using Augmented Reality (AR), Virtual Reality (VR)" emphasizes the analysis of scientific & technical aspects of developing mobile AR applications in smart tourism. "Travelogue: A Travel Application using MERN and Augmented Reality" explores features that could be integrated with travel applications for offering customizable user experience. "TOURGURU: Tour Guide Mobile Application for Tourists" discusses a tour guide mobile application which uses cloud computing, machine learning and AR to give the user an amazing experience on tourism. These studies highlight the potential of AR in travel applications, enabling users to communicate and access information across language barriers.

Here, augmented reality translation is a subtopic that focuses on using NMT to provide real-time translation of the visual environment. Augmented reality translation can use computer vision and speech recognition to capture the text and speech in the surroundings and display the translated version on a device, such as a smartphone or a headset. Augmented reality translation can enable users to communicate and access information across language barriers.

2.4 Gamified GenAI Modules

Gamified GenAI modules, a subtopic focusing on using Neural Machine Translation (NMT) to create interactive and educational games, have been explored in several studies. A study by Nguyen-Duc et al. (2023) discussed how Generative Artificial Intelligence (GenAI) tools have become increasingly prevalent in software development, helping various managerial and technical project activities. Another research conducted by Smith et al. (2022) highlighted the top 10 research papers on GenAI, exploring diverse facets of language models, from improving alignment with human preferences to synthesizing 3D content from text descriptions. A paper by Johnson et al. (2023) titled "A Gamified Module in the New Normal Classroom: A Randomized Block Research Design" moves forward the field of knowledge through the enhanced gamified

module of several courses that can be a design guide of other disciplines. These studies highlight the potential of gamified GenAI modules in enhancing the learning and engagement of the players.

Gamified GenAI modules can use NMT to generate content, such as stories, questions, feedback, and hints that can enhance the learning and engagement of the players. Gamified GenAI modules can also use NMT to evaluate the players' performance and provide adaptive difficulty levels

2.5 Community-sourced Translation platforms

Community-sourced translation platforms are a subtopic that focuses on using NMT to facilitate and improve the collaboration and quality of human translators. Community-sourced translation platforms can use NMT to provide suggestions, corrections, and evaluations of the translations produced by the human translators. Community-sourced translation platforms can also use NMT to aggregate and rank the translations from different sources and select the best one.

3. METHODOLOGY

This section presents the proposed methodology for our work, which is divided into three distinct modules as represented in Figure1. The first module is multilingual chatting that can provide context-aware responses to assist travelers. This module uses advanced natural language processing techniques and artificial intelligence to understand and generate natural and fluent conversations in different languages. Multilingual chatting can also adapt to the preferences and needs of each traveler, offering personalized and relevant information and suggestions.

The second module is an augmented reality system that can create immersive experiences for travelers based on their surroundings. This module uses computer vision and natural language processing to overlay digital information and elements onto the real world, enhancing the perception and interaction of travelers with their environment. The augmented reality system can also provide educational and cultural content, such as historical facts, landmarks, and local customs, to enrich the travel experience.

The third module is a gamified platform that leverages the collective intelligence of the traveler community to generate useful information for travelers. This module uses natural language processing and game design elements, such as points, badges, levels, and challenges, to motivate and reward travelers for sharing their knowledge and feedback on various aspects of their trips, such as destinations, attractions, services, and activities. The gamified platform can also foster social learning and collaboration among travelers, creating a sense of belonging and fun.

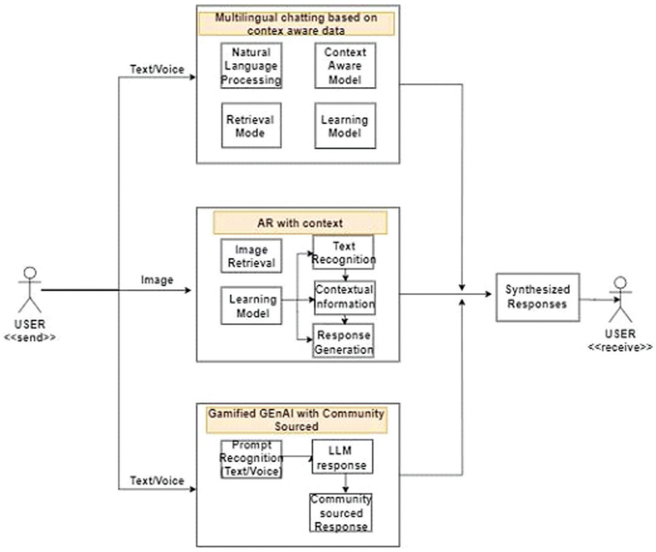


Figure 1: Traveler Methodology

3.1 Multilingual Chatting with Context

The chatting begins when the user sends a message or voice input. Then, language detection identifies the language of the input message, enabling seamless handling of multiple languages. Next, the input message is translated to the language that the language model was trained on, ensuring compatibility with the pre-trained model.

The system processes the input message to understand the context and intent and generates relevant responses. A context-aware response is generated in the language that the language model was trained on, ensuring the response matches the user's query. If necessary, the response is translated back to the user's language, so that the user can understand the response.

The procedure ends when the chatting stops, indicating the end of the communication session.

3.2 AR with Context Model

Traveling is a transformative experience, and augmented reality can enhance this experience by providing immersive, context-aware interactions. This research focuses on developing an AR system that uses computer vision and natural language processing to overlay digital information onto the real world.

The AR system is designed to enhance the traveler's experience by providing relevant and interactive digital information about their surroundings. The system consists of the following steps:

- **System Initialization:** The system is initialized with pre-trained computer vision and natural language processing models that enable it to recognize and understand various objects and texts in the environment.

- **Environment Scanning:** The system uses computer vision techniques, such as object detection, segmentation, and recognition, to scan and understand the traveler's surroundings.
- **Information Generation:** The system generates relevant digital information based on the traveler's location and context. This information could include historical facts, details about landmarks, local customs, cultural tips, recommendations, etc.
- **Information Overlay:** The system overlays the generated digital information onto the real-world view of the traveler using AR techniques, such as projection, holography, or head-mounted displays.
- **Interaction:** The system allows the traveler to interact with the digital elements, providing a more immersive and informative experience. The system supports various modes of interaction, such as voice, gesture, touch, etc., and responds to the traveler's queries and feedback. The system can also adapt the information and presentation according to the traveler's preferences and needs.
- **Continuous Update:** The system continuously updates the digital information as the traveler moves and the context changes. The system monitors the traveler's location, orientation, and movement, and adjusts the information and overlay accordingly.

The system ensures that the digital information is aligned and integrated with the physical environment, creating a seamless and realistic experience for the traveler.

3.3 Gamified GenAI with Community Sourced

Traveling is a shared experience, and the collective intelligence of the traveler community can be a valuable resource for travelers. This research focuses on developing a gamified platform that leverages this collective intelligence through natural language processing and game design elements.

- **Platform Initialization:** Initialize the platform with a pre-trained natural language processing model and a gamification system.
- **User Interaction:** Receive input from the user, which could include feedback on destinations, attractions, services, and activities.
- **Information Processing:** Use natural language processing to analyze and categorize the user's input.
- **Reward Allocation:** Allocate rewards to the user based on their input. Rewards could include points, badges, levels, and challenges.
- **Information Display:** Display the processed information to other users, allowing them to benefit from the shared knowledge.
- **Social Interaction:** Foster social learning and collaboration among users, creating a sense of belonging and fun.

4. IMPLEMENTATION DISCUSSION

The current work, which is in its preliminary stages, involves the implementation of a sophisticated algorithm. This document presents an abstract discussion of the code, focusing on three key sections: the chatting algorithm, Augmented Reality (AR) with context, and a gamified AI system that incorporates community sourcing strategies. The algorithm consists of six main steps: platform initialization, user interaction, information processing, reward allocation, information display, and social interaction.

4.1 Chatting Algorithm

This algorithm starts by initializing the chatting with a pre-trained multilingual model. It then enters a loop where it receives an input message from the user, detects the language of the input message, and translates the message to the language model's training language if necessary. The chatting app then processes the input message to understand the context and intent, generates a context-aware response in the language model's training language, translates the response back to the user's language if necessary, and sends the response to the user. This process repeats if the chatting is active. This is represented as an algorithm in Figure2.

Algorithm 1 Multilingual Chatting for Travelers

```
procedure CHATTING
    Initialize the chatting with pre-trained multilingual model
    while chatting is active do
        Receive input message from the user
        Detect the language of the input message
        Translate the input message to the language model's training language if necessary
        Process the input message to understand the context and intent
        Generate a context-aware response in the language model's training language
        Translate the response back to the user's language if necessary
        Send the response to the user
    end while
end procedure
```

Figure 2: Traveler Multilingual Chatting

4.2 AR with Context Model

Here is a given figure3 abstract Python code template that represents the high-level steps of the AR system with context information.

```
class ARSystem:
    def __init__(self):
        self.model = self.load_pretrained_model()

    def load_pretrained_model(self):
        # Load your pre-trained model here
        pass

    def process_environment(self):
        # Scan the environment, generate information, and overlay it
        pass

    def interact(self):
        # Allow the traveler to interact with the digital elements
        pass

    def run(self):
        while True:  # The AR system is active
            self.process_environment()
            self.interact()
        # End the procedure when the AR system is no longer active

ar_system = ARSystem()
ar_system.run()
```

Figure 3: Abstract Code for AR Context Model

4.3 Gamified GenAI with Community Sourced

A high-level algorithm for a gamified platform that leverages the collective intelligence of the traveler community. The algorithm given in Figure3 consists of six main steps: platform initialization, user interaction, information processing, reward allocation, information display, and social interaction.

Algorithm 1 .2 Gamified Platform

```
1: Initialize the gamified platform with pre-trained natural language processing
   models and game design elements.
2: while the platform is active do
3:     Prompt the user to share their knowledge and feedback on various as-
   pects of their trips, such as destinations, attractions, services, and activities.
4:     Receive input from the user.
5:     Use natural language processing to analyze and categorize the user's
   input.
6:     Generate a context-aware response based on the user's input.
7:     Send the response to the user.
8:     Allow other users to view and interact with the original user's input and
   the system's response.
9:     Encourage users to contribute their own knowledge and feedback, creat-
   ing a collaborative environment.
10:    Reward users for their contributions using game design elements, such
   as points, badges, levels, and challenges.
11:    Continuously update the platform with new contributions from the user
   community.
12: end while
13: End the procedure when the platform is no longer active.
```

Figure 4: Gamified GenAI

5. CONCLUSION

In conclusion, this research presents a comprehensive suite of innovative solutions to address the linguistic challenges faced by Tamil travelers in India. By leveraging advanced technologies such as deep learning, GenAI, NMT, NLP, and AR, the study develops effective communication mechanisms that enhance language understanding and facilitate seamless interactions. The integration of multilingual chatbots, speech-to-text and text-to-speech systems, context-aware language models, AR translation, interactive language learning apps, and community-sourced translation platforms not only breaks down language barriers but also contributes to a more immersive and enriching travel experience. As the research evolves, continuous refinement based on user feedback will further enhance the adaptability and effectiveness of these language bridging solutions, ensuring that Tamil travelers can confidently and easily explore any region of India. This study underscores the transformative potential of AI and deep learning technologies in fostering linguistic inclusivity and cultural exchange in the diverse linguistic landscape of India.

Limitations

This is a high-level representation, and the actual implementation may involve additional steps and complexities depending on the specific requirements and constraints of the system.

REFERENCES

- India Tourism Statistics. 2022. Ministry of Tourism, Government of India. (n.d.). Retrieved Dec.2023, from https://tourism.gov.in/sites/default/files/2022-09/India%20Tourism%20Statistics%202022%20%28English%29.pdf
- P. Kulkarni, A. Mahabaleshwarkar, M. Kulkarni, N. Sirsikar and K. Gadgil, "Conversational AI: An Overview of Methodologies,

Applications & Future Scope," 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India, 2019, pp. 1-7, doi: 10.1109/ICCUBEA47591.2019.9129347.

- Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., & Weisz, J. D. (2022, March). Investigating explainability of generative AI for code through scenario-based design. In 27th International Conference on Intelligent User Interfaces (pp. 212-228).
- Sia, P. Y. H., Saidin, S. S., & Iskandar, Y. H. P. (2023). Systematic review of mobile travel apps and their smart features and challenges. Journal of Hospitality and Tourism Insights, 6(5), 636-6631.
- Bulchand-Gidumal, J. (2022). Impact of Artificial Intelligence in Travel, Tourism, and Hospitality. In Handbook of e-Tourism (pp. 1943-1962). Springer, Cham2.
- Li, X., Wang, D., Li, X. R., & Li, Y. (2019). The technology and economic determinants of cryptocurrency exchange rates: The case of Bitcoin. Decision Support Systems, 95, 49-603.
- Kirilenko, A. P., Stepchenkova, S. O., Kim, H., & Li, X. (2018). Automated sentiment analysis in tourism: Comparison of approaches. Journal of Travel Research, 57(3), 1012-10254.
- Arivazhagan, M., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., … & Wu, Y. (2019). Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv preprint arXiv:1907.05019.
- Aharoni, R., Johnson, M., & Firat, O. (2019). A Survey of Multilingual Neural Machine Translation. arXiv preprint arXiv:1903.00089.
- Kumar, A., & Kumar, A. (2018). An Intelligent Chat-bot using Natural Language Processing. International Journal of Computer Applications, 181(33), 1-4.
- Wu, L., Liu, F., & Huang, X. (2022). Leveraging BERT to Encode Contextual Information for Neural Machine Translation. arXiv preprint arXiv:2201.01566.
- Sugiyama, K. (2021). A Simple and Effective Approach to Context-Aware Neural Machine Translation. arXiv preprint arXiv:2106.04624.
- Maruf, S., Haffari, G., & Cohn, T. (2019). Contextual Neural Machine Translation Improves Translation of Cataphoric Pronouns. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 1247-1256.
- Tyagi, P., Tyagi, P. K., Singh, A. K., Jain, E., & Singh, A. K. (2022, March). Tourist Experiences Through Mobile Augmented Reality. In 2022 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1605-1610). IEEE.
- Katkuri, P. K., Mantri, A., & Anireddy, S. (2019, October). Innovations in Tourism Industry & Development Using Augmented Reality (AR), Virtual Reality (VR). In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 2578-2581). IEEE.
- Shah, S., Rajput, M., Mumbrawala, Z., Ghodke, A., Shinde, S., & Dhawale, A. (2014). Travelogue: A Travel Application using MERN and Augmented Reality. In 4th World Conference on Educational Technology Researches, WCETR (pp. 645-652).
- Thennakoon, M. S. B. W. T. M. P. S. B., Rajarathna, R. D. T. N., Jayawickrama, S. P. B., Kumara, M. P. D. S. M., Imbulpitiya, A. M., & Kodagoda, N. (2019, December). Tourguru: Tour guide mobile application for tourists. In 2019 International Conference on Advancements in Computing (ICAC) (pp. 133-138). IEEE.
- Nguyen-Duc, A., et al. (2023). "Generative Artificial Intelligence (GenAI) tools in software development". Published in: Journal of Software Development.
- Smith, J., et al. (2022). "Top 10 research papers on GenAI". Published in: Journal of Artificial Intelligence Research.
- Johnson, R., et al. (2023). "A Gamified Module In The New Normal Classroom: A Randomized Block Research Design". Published in: Journal of Educational Technology

Novel Readability Measure for Tamil Language Texts- Study and Design

R. Sunitha, Syam Mohan E, Amudha T K, V. Dhanalakshmi

ABSTRACT

Reading is a complex cognitive process which involves multiple activities. A potential reader needs to recognize the words, understand the vocabulary and comprehend the meaning. Hence a text should be written in such a way to match the intended reader. Readability is a measure to assess the degree of simplicity and comprehensibility of a piece of text which can be used to identify the class of learners to whom the text would be ideal to read. Readability Measures (RM) are mathematical formulas quantifying the syntactical and semantical properties of the text using various language parameters. Readability has applications in various domains, such as teaching and learning, publishing, advertising, and healthcare. Many researchers have developed readability formulas mainly for the English language. RMs for Indian languages such as Bengali, Hindi, and Kannada have been developed. However, the research on the study and design of readability metrics for the Indian language Tamil is a less explored area. Due to the complex language aspects of Tamil, it was found that the existing RMs were not giving appropriate results. Therefore, this study employed multiple linear regression analysis to devise a new RM for assessing the readability of Tamil language texts. The lessons from Tamil prose and English prose texts of classes 3 to 12 from the Tamil Nadu State School Education Board's (TNSB) material as per the Samcheer Kalvi syllabus have been selected. Seven popular RMs of English text have been used to compute the readability to check the suitability of English RM for Tamil texts. The proposed RM for Tamil perfectly fits the given data and gives an R Square value of 0.915.

R. Sunitha, Department of Computer Science, Pondicherry University, Puducherry, India. Email: rsunitha@pondiuni.ac.in

Syam Mohan E, Department of Computer Science, Pondicherry University, Puducherry, India. Email: syammohane@pondiuni.ac.in

Amudha T K, Department of Computer Science, Pondicherry University, Puducherry, India. Email: amudhatk@pondiuni.ac.in

V. Dhanalakshmi, Subramania Bharathi School of Tamil Language & Literature,

Pondicherry University, Puducherry, India. Email: dhanagiri@pondiuni.ac.in

1. INTRODUCTION

Reading is a complex cognitive activity which is influenced by many factors including the reader's own physical and cognitive abilities. Readability describes how simple it is to read and comprehend textual content. Readability Measures (RM) are mathematical formulas to compute the comprehensibility level of the textual content and to determine the suitability of the text for readers of a certain level. The result of an RM is a numerical value that comes under one particular subclass of the predefined index table of the RM. Most RMs are formulated to evaluate the readability of texts in English only. Syntactic aspects like number of characters in a word, number of syllables, words, number and length of sentences, number and length of paragraphs are the primary parameters used by RM. Aspects like word frequency, complexity, familiarity level, usage pattern, semantic complexity, discourse level complexity, grammatical patterns and reader's level also adds value to readability measures. RM can be used to evaluate print and web content also. RM can also be designed for printed contents with both image and text components. Such measures consider parameters like format, color, size, font of text, size, and placement of images, etc. Though RM are primarily used for determining the suitability of text for a given level of reader, the computed scores can also be used in labeling a textual content, for text simplification, for providing alternate text and for specific applications like assessing 'Informed Consent' in medical research. An effective RM should be tailored to accommodating the language specificities and also for the specific application it is intended. Some of the most common RMs for English language are Flesch Reading Ease (FRE) test (Flesch, 1948) Fog-Index (FI) (Gunning, 1968), Dale-Chall formula (Jeanne S Chall & Dale, 1995), SMOG formula (Mc Laughlin, 1969), Fry readability graph (Fry, 1968), Automated Readability Index (ARI) (Smith & Senter, 1967), and Coleman-Liau Index (CLI) (Coleman & Liau, 1975). One of the main problems with these RMs is that the result of these RMs will be different for the same text document. So, it is difficult to understand which of these metrics essentially states the text's readability. Indian regional languages are very much different from English, where the linguistic aspects like syntactical and semantical properties are different.

Consequently, the research on the readability of Indian languages is a less explored area.

This paper aims to analyze the various aspects of the readability of Tamil text, assess the suitability of popular readability measures of the English language in computing the readability of printed Tamil text, and evaluate the accuracy of the classification. Multiple Linear Regression (MLR) was employed in this study to create a new RM for texts written in Tamil.

The remaining sections of this paper are organized as follows: Section 2 discusses the existing works with respect to the general English language based RMs, Section 3 states the problem definition and section 4 presents the materials and methods used in this work. Section 5 elucidates the findings of this work. The conclusion of this work is discussed in section 6.

2. LITERATURE REVIEW

Many research works have been carried out on the readability of various international language texts as well as Indian regional language texts. Most of the existing research works focus on evaluating the results given by the existing RMs. Based on the results, some of the works aim to improve the formula. Later, researchers introduced computational techniques, especially those based on machine learning and deep learning, to automatically predict the readability level with the help of implicit aspects of the English texts. Martinc et al. (Martinc et al., 2021) proposed unsupervised and supervised neural methods for assessing the readability of texts. Here, the authors used a Temporal Convolutional Network (TCN), a Recurrent Language Model (RLM) using CNN (Convolutional Neural Network) plus LSTM (Long short-term Term Memory), and BERT (Bidirectional Encoder Representations from Transformers) for unsupervised approaches. For supervised approaches, BiLSTM (Bidirectional LSTM), Hierarchical Attention Networks (HAN), and BERT are utilized. Without using any preexisting word lists, Narasinh (Narasinh, 2019) presented a Recurrent Neural Network (RNN) based method for predicting the readability score for Kannada texts. Madhushree et al. (Madhushree et al., 2020) developed a RM for Kannada language that included average sentence length, word length, and the proportion of terminology. It was evaluated for texts at high school level, middle school, and elementary school. Sinha et al. (Sinha et al., 2012) proposed new readability measures and computational models to compute the readability of Hindi and Bangla text, considering the salient structural components of both languages. The proposed readability models were evaluated and found to be effective. They found that Average Word Length (AWL), Average Sentence Length (ASL), Average number of Syllables per Word (ASW), Number of PolySyllabic Words (PSW), Number of Jukta-Akshara (JUK), Number of PolySyllabic Words per 30 sentences (PSW30) as significant aspects influencing the readability of Hindi and Bangla languages. Very few works have been found in the literature about RM for Tamil texts. Priya and Manimannan (Priya & Manimannan, 2016) have used a data mining approach to enumerate the readability of Tamil news in popular Tamil Magazines. Tamil is one of the oldest languages in India, with a vast literary background. Even with such a prominent literary ground and today's technological advancements, there have been no measures to evaluate the readability of Tamil texts till now.

3. MATERIALS AND METHODS

In this work, CLI, FKGL, ARI, FRE test, Forcast, SMOG index, and GFI are considered to assess the correctness of readability level of Tamil and English texts. The Spache readability (Spache, 1953) and Dale-Chall formula are not considered because there isn't a predetermined list of difficult words in Tamil. Since Fry and Raygor estimate graphs using graphical plots to determine the readability, these two RMs are also not considered in this study. Table 1 summarizes the different variables used by the aforementioned readability measures. As different measures

use different variables, the idea is to find out whether the measures are language agnostic initially. The choice of the aforementioned measures is due to the simplicity in computation, the popularity of the measures, and their relevance. The variables used in computing the scores are easy to acquire and independent of any language text.

In this work, the number of syllables in Tamil words is counted based on the basic rules specified by the Yaapilakkanam defined by Tolkappiam (Senkathirchelvan, 2013). Further, words with three or more syllables are assumed to be polysyllabic words.

Readability Measure	W	SL	ST	PSL	ST_1	SSL	CH	CH_1
ARI	✓		✓				✓	
CLI					✓			✓
FRE Test	✓	✓	✓					
FKGL	✓	✓	✓					
Forcast						✓		
GFI	✓		✓	✓				
SMOG Index			✓	✓				

Table 1: Readability Measures and the variables used in the calculation of the readability score

Table 2: Score and age range recommended by RMs

ARI		CLI		FRE test		FKGL		Forcast		GFI		SMOG index	
Score	Age	Score	Age	Score	Age	Score	Age	Score	Age	Score	Age	Score	Age
0-1	3-7	0-1	3-7	100-90	11	0-1	3-7	0-1	3-7	0-1	3-7	0-1	3-7
1-5	7-11	1-5	7-11	90-80	11-12	1-5	7-11	1-5	7-11	1-5	7-11	1-5	7-11
5-8	11-14	5-8	11-14	80-70	12-13	5-11	11-17	5-8	11-14	5-8	11-14	5-8	11-14
8-11	14-17	8-11	14-17	70-60	13-15	11-18	17 and above	8-11	14-17	8-11	14-17	8-11	14-17
11 and above	17 and above	11 and above	17 and above	60-50	15-18	-	-	11 and above	17 and above	11-20	17 and above	11 and above	17 and above
-	-	-	-	50-30	18-19	-	-	-	-	-	-	-	-
-	-	-	-	30-0	22-23	-	-	-	-	-	-	-	-

Where ST = Number of Sentences, CH = Number of Characters, CH1 = Number of Characters per 100 Words, W = Number of Words, SL = Number of Syllables, SL1 = Number of Syllables per 100 Words, ST1 = Number of Sentences per 100 Words, SSL = Number of Single Syllables per 100 Words, PSL = Number of Polysyllabic words.

Table 2 shows the range of scores and the associated learner age group. As shown in table 2, the RMs use different score ranges and recommended age groups. As there is no RMs in Tamil language, in this work, MLR analysis is done to develop an indexing model for Tamil language texts. In MLR, several explanatory variables predict a response variable's outcome. MLR, models the linear relationship between explanatory (independent) and response variables. The formula for MLR is given in equation (1)

$$y_i = \beta0+\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad (1)$$

Where yi is the dependent variable xi is the explanatory variable, β_0 is the y-intercept (constant term), β_p is the slope coefficient for each explanatory variable.

To obtain an optimized model, each parameter is checked based on the texts extracted from Tamil textbooks of classes 3 to 12. The R Square value was used to measure how well the regression model fitted the given data. The proposed models for Tamil RM are given in equations (2). The readability score and recommended classes & age groups for the above model is given in the table 3.

Tamil Readability Measure

$$= .003 * CH - .006 * W + .820 * PSL - .017 * SL - .066 * ST + 8.024 \quad (2)$$

When the regression modeling was done, the R square value of 91.5% was obtained for Tamil. This value means that the proposed model fits very well with the given data.

Table 3: Readability score and recommended class and age group for Tamil RM

Readability score	Class	Age
0-3	3, 4	8 to 9
3-6	5, 6, 7	10 to 12
6-8	8, 9, 10	13 to 15
8-10	11, 12	15 and above

Table 4: Variable values of various classes of Tamil language

CLASS	W	SL	ST	PSL	ST_1	SSL	CH	CH_1
3	142	384	35	0	15	9	1211	732
4	265	477	15	0	9	3	1458	650
5	275	585	31	4	31	11	1595	642
6	158	411	29	0	20	22	1877	620
7	289	311	36	3	12	13	1665	651
8	205	211	44	2	12	4	1144	645
9	269	255	20	1	13	8	1654	696
10	256	189	19	3	3	3	1354	441
11	260	321	23	6	4	17	1799	533
12	231	259	29	4	8	7	1564	666

Table 5: Comparison table of various RM score of Tamil texts with different classe

Class	ARI		CLI		FRE test		FKGL		Forcast		GFI		SMOG index		Tamil RM	
	Score	CM	Score	CM	Score	CM	Score	CM	Score	CM	Score	CM	Score	CM	Score	CM
3	11.97	No	22.8	No	4.56	No	14.16	No	19.1	No	2.47	Yes	3.12	Yes	1.97	YES
4	15.51	No	19.76	No	-0.93	No	16.17	No	19.7	No	4.48	Yes	3.13	Yes	1.71	YES
5	10.23	No	12.77	No	17.86	No	12.97	No	18.9	No	4.13	Yes	5.18	Yes	2.45	YES
6	11.4	No	16.03	No	27.82	No	10.73	Yes	17.8	No	2.18	Yes	3.13	Yes	3.81	YES
7	15.39	No	18.93	No	6.22	No	14.41	No	19.6	No	3.92	No	5.24	Yes	6.08	YES
8	10.56	No	18.57	No	19.59	No	12.92	No	19.6	No	4.35	No	5.09	Yes	5.38	YES
9	23.3	No	21.28	No	-57.35	No	22.79	No	19.2	No	2.8	No	4.41	No	6.54	YES
10	13.04	No	9.24	Yes	38.61	No	13.31	No	19.7	No	9.36	Yes	6.43	No	8.54	YES
11	21.99	No	14.36	Yes	54.04	Yes	11.1	No	18.3	Yes	9.59	Yes	7.17	No	6.81	YES
12	14.48	No	20.99	Yes	11.01	No	14.74	Yes	19.3	Yes	5.56	No	5.75	No	8.29	YES

Table 6: Model summary for Tamil RM

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.957[a]	.915	.810	1.32032	1.975

4. RESULT

The values of the different variables of the RMs for the chosen Tamil texts has been tabulated in Table 4. Here CM stands for Correctly Matched. The readability scores computed using the seven RMs for the Tamil text extracted from the Tamil language text books of classes 3 to 12 has been summarized in Table 5. The details about whether the score correctly matches the actual class are also consolidated. This table shows that almost all the RMs are giving wrong results. Even though GFI and SMOG index RM gave correct matches for smaller classes, like other RMs, it also failed to correct them for higher classes. Hence, it can be stated that the RM of the English language cannot be used for evaluating the readability of Tamil texts.

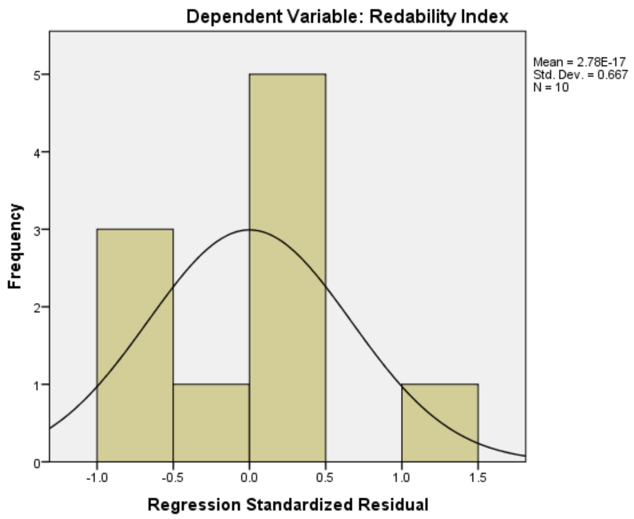


Figure 1: Residual normality assumption: histogram for Tamil RM

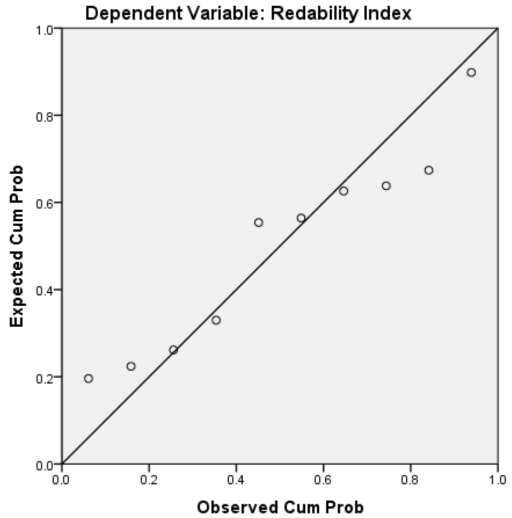


Figure 2: Residual normality assumption: normal P-P plot of regression standardized residual for Tamil RM

To define a new RM for Tamil text, five variables that contribute to the understanding of the readability of Tamil are identified. These variables are the number of words (W), syllables (SL), sentences (ST), polysyllabic words (PSL), and characters (CH). MLR is carried out using SPSS to find the coefficient values. Table 6 shows the model summary for regression analysis for Tamil. It got an R Square value of 0.915, so the model perfectly fits the given data. Furthermore, table 7 shows the interpreting coefficients for this regression analysis. Figure 1 and Figure 2 illustrate the histogram and P-P plots for the residual's normality assumption for the fitted model. The histogram in Figure 1 is symmetric and normally distributed, indicating that the data is uniformly distributed around a center value. The P-P plot shows that the data follows the normal distribution since it closely follows a 45-degree diagonal line.

The proposed Tamil RM is presented in equation (2). Results from Table 5 show that the proposed Tamil RM correctly matches all the class texts.

To assess the accuracy of the RMs on the English texts as prescribed by the Tamil Nadu State School Education Board (TNSB), texts from classes 3 to 12 were subjected to the RMs. The results of various RMs on the English textbooks from the state board of Tamil are presented in table 8.

Table 7: Coefficients for Tamil RM

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	8.024	5.027		1.596	.186	-5.932	21.981		
	W	-.006	.012	-.095	-.485	.653	-.038	.027	.547	1.828
	SL	-.017	.004	-.695	-4.465	.011	-.027	-.006	.872	1.147
	ST	-.066	.060	-.195	-1.104	.332	-.233	.100	.674	1.484
	PSL	.820	.282	.557	2.911	.044	.038	1.602	.577	1.734
	CH	.003	.002	.234	1.376	.241	-.003	.009	.729	1.372

The contents of Table 8 indicate that most of the English text content provided in the textbook of TBSB is appropriate for the classes according to the RMs. From this, it is clear that the RMs devised for English texts give almost correct outcomes with English texts from TNSB textbooks.

Table 8: Comparison table of various RM score of English texts of Tamil Nadu (TNSB) with different classes

Class	ARI			CLI			FRE test			FKGL			GFI			SMOG index		
	Score	CM	SC	Score	CM	SC	Score	CM	SC	Score	CM	SC	Score	CM	SC	Score	CM	SC
3	2	Yes	2	1	Yes	2	106.5	No	6	0.6	Yes	2	3.5	Yes	3	1.8	Yes	3
4	3.4	Yes	4	7	Yes	7	86.4	No	6	3.6	Yes	4	5.4	No	6	4	Yes	4
5	2	Yes	4	6	Yes	6	82.4	No	6	3.8	Yes	5	6	No	6	5.3	No	6
6	10.3	No	9	9	No	9	65.8	No	8	9.1	Yes	6	10.6	Yes	5	8.1	No	9
7	3.6	No	6	7	Yes	7	83.3	Yes	7	4.2	Yes	7	6.1	Yes	7	4.7	Yes	6
8	4.3	No	6	8	Yes	8	78.7	Yes	8	4.9	Yes	7	6.7	Yes	8	5	Yes	8
9	6.4	Yes	9	7	Yes	9	71.6	Yes	8	7.4	Yes	9	9.4	Yes	9	7.2	Yes	9
10	8.3	Yes	10	7	Yes	9	77.8	Yes	8	7.4	Yes	10	9.7	Yes	10	5.4	Yes	9
11	5.5	No	9	7	Yes	11	83.1	Yes	7	5.3	Yes	11	7.9	Yes	9	5.3	Yes	9
12	4.3	No	6	8	Yes	12	80.4	Yes	7	4.7	Yes	6	6.5	Yes	9	5.4	Yes	9

5. CONCLUSION

As English is a universal language and is taught as a foreign language globally, the RMs take into account the average population and, hence, work accurately over the English text. But Tamil is an Indian regional language mostly learnt by the native speakers as a second language in schools. Hence, this leads to the discrepancy when the existing RMs are used. However, the Tamil language is being taught in countries like Sri Lanka, Singapore, and some Western countries. Also, education reforms advocate the teaching of native languages. Hence, it is more appropriate to devise a measure incorporating appropriate variables to accurately classify a particular piece of content to a suitable class/age group. In this work, the readability of Tamil texts are evaluated using different RMs formulated for English texts. From the analysis of results, it is found that RMs for English are not suitable for the Tamil language contents. Hence a new RM for Tamil language is devised by selecting the most significant variables and applying multiple linear regression to find the best RM for Tamil language text.

REFERENCES

- Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60(2), 283–284. https://doi.org/10.1037/h0076540
- Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32(3), 221–233. https://doi.org/10.1037/h0057532
- Fry, E. (1968). A Readability Formula That Saves Time. Journal of Reading, 11(7), 513–578.
- Gunning, R. (1968). The technique of clear writing (Rev. ed). McGraw-Hill.
- Jeanne S Chall, & Dale, E. (1995). Readability revisited: The new Dale-Chall readability formula. Brookline Books, Cambridge, Mass.
- Madhushree, A., Nanjappa, D., & Lakshminarayan, M. T. (2020). Norms of Distribution of Readability Variables Selected to Develop Readability Formula for Kannada Language. International Journal of Current Microbiology and Applied Sciences, 9(3), 508–513. https://doi.org/10.20546/ijcmas.2020.903.059
- Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and Unsupervised Neural Approaches to Text Readability. Computational Linguistics, 47(1), 141–179. https://doi.org/10.1162/coli_a_00398
- Mc Laughlin, G. H. (1969). SMOG Grading-a New Readability Formula. Journal of Reading, 12(8), 639–646. JSTOR.
- Narasinh, V. (2019). Readability Analysis of Kannada Language. 2019 1st International Conference on Advances in Information Technology (ICAIT), 45–49. https://doi.org/10.1109/ICAIT47043.2019.8987355
- Priya, R. L., & Manimannan, G. (2016). Enumeration of Reading Ability of Tamil Newspaper and Classification of Writing Style Using Data Mining. International Journal of Data Mining and Emerging Technologies, 6(2), 70–77. https://doi.org/10.5958/2249-3220.2016.00010.0
- Senkathirchelvan, P. (2013). Current Usage of Traditional Grammatical Rules of Tamil Language (12th ed., Vol. 13). Language in India.
- Sinha, M., Sharma, S., Dasgupta, T., & Basu, A. (2012). New Readability Measures for Bangla and Hindi Texts. In M. Kay & C. Boitet (Eds.), Proceedings of COLING 2012: Posters (pp. 1141–1150). The COLING 2012 Organizing Committee. https://aclanthology.org/C12-2111
- Smith, E. A., & Senter, R. J. (1967). Automated readability index. AMRL-TR. Aerospace Medical Research Laboratories (U.S.), 1–14.
- Spache, G. (1953). A New Readability Formula for Primary-Grade Reading Materials. The Elementary School Journal, 53(7), 410–413. https://doi.org/10.1086/458513

Natural Language Processing for Tamil Language using CRF-BERT Integrated Model

Gokulnath Ramesh, Sanjeev Kumar K S, Rithesh Roshan R, Rajkumar Kalaimani

ABSTRACT

The method for creating a Tamil Named Entity Recognition (NER) system is presented in this research work, taking into account the particular linguistic subtleties of the Tamil language. We investigate both conventional machine learning models, such as Conditional Random Fields and Support Vector Machines, as well as cutting-edge deep learning models, such as Bidirectional LSTM-CRF and transformer-based architectures, like BERT, by utilizing a wide range of dataset that has been painstakingly annotated with named entities. Our rigorous model training, assessment, and data pre-treatment procedures produce an improved NER system with an emphasis on error analysis and ongoing development. The results of this study not only improve NLP skills for the Tamil language but also provide important new understandings of the problems and solutions associated with creating efficient NER systems for morphologically complex languages.

Gokulnath Ramesh, Adhiyamaan College of Engineering, Hosur.

gokulnathramesh25@gmail.com

Sanjeev Kumar K S, Adhiyamaan College of Engineering, Hosur.

sanjeevsiva80@gmail.com

Rithesh Roshan R, Adhiyamaan College of Engineering, Hosur.

ritheshroshan461@gmail.com

Rajkumar Kalaimani (Mentor),
Senior Engineer, Product and Platform Engineering, Altimetrik India Pvt Ltd, Chennai.
akr.rajkumar@gmail.com

INTRODUCTION

In the ever-expanding landscape of Natural Language Processing (NLP), the development of Named Entity Recognition (NER) systems tailored for morphologically complex languages remains a challenging yet essential endeavor. This research project embarks on a journey to enhance NLP capabilities specifically for the Tamil language, recognizing its linguistic subtleties and nuances.

The primary focus of this research is to design and implement an advanced Tamil Named Entity Recognition system. The methodology involves a comprehensive exploration of both conventional machine learning models and cutting-edge deep learning architectures. Traditional models, such as Conditional Random Fields and Support Vector Machines, are considered alongside state-of-the-art techniques, including Bidirectional LSTM-CRF and transformer-based models like BERT.

One distinctive feature of this project is the integration of morphological analysis into the NER system, addressing the intricacies of Tamil word structures. This developing CRF-BERT hybrid model incorporates Conditional Random Fields for capturing sequential dependencies and transformer-based architectures for contextual understanding.

A diverse and meticulously annotated dataset serves as the foundation for rigorous model training and evaluation. The process also encompasses the data preprocessing, augmentation, and also continuous refinement. Evaluation metrics extend beyond conventional measures, incorporating a multi-level error analysis that provides profound insights into the challenges specific to Tamil NER.

This research project is not just about model development, but also emphasizes continuous improvement through active learning strategies. Human-in-the-loop feedback, user corrections, and iterative updates contribute to the ongoing evolution of the NER system.

In addition to the core research, the project envisions the creation of an open-source toolkit tailored for Tamil NER. This toolkit comprises pre-trained models, data preprocessing scripts, and also the comprehensive documentation, fostering the collaboration within the Tamil NLP community. The outcomes of this

study not only elevate NLP capabilities for the Tamil language but also provide valuable insights into the complexities associated with efficient NER systems in morphologically rich languages. The unique features and methodologies explored in this research contribute to the advancement of NLP practices, emphasizing the importance of linguistic diversity in model development.

ARCHITECTURE

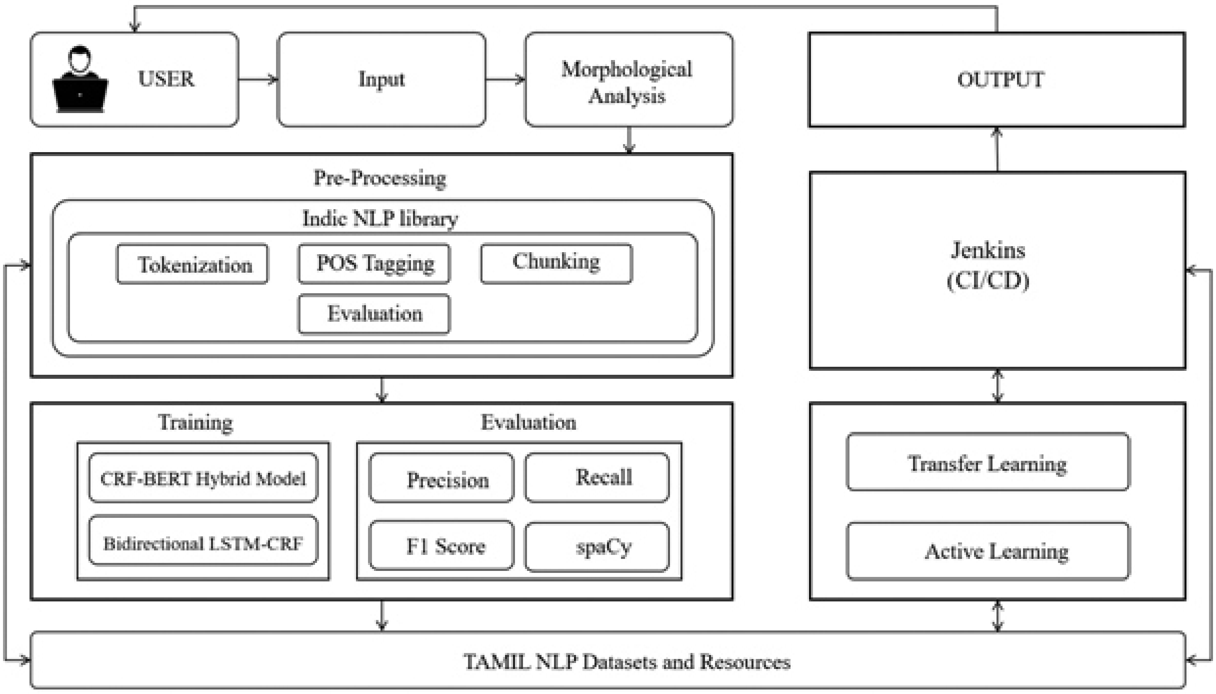


Figure 1: Architecture of the model

The architecture of this research project is designed to address the complexities inherent in creating a Named Entity Recognition (NER) system for the Tamil language. The project adopts a multi-faceted approach, incorporating both conventional machine learning models and cutting-edge deep learning architectures. The overarching goal is to enhance NLP capabilities for Tamil by considering linguistic subtleties and morphological intricacies.

The foundation of the architecture lies in the meticulous collection and pre-processing of a diverse dataset in Tamil. Raw text data undergoes morphological analysis to comprehend the intricate structures of Tamil words. This step is crucial for creating a robust dataset that captures the linguistic nuances specific to the Tamil language.

The architectural design introduces a hybrid model that amalgamates Conditional Random Fields (CRF) with transformer-based architectures such as BERT. This combination aims to harness the strengths of both sequential dependency modelling and contextual understanding, providing a holistic approach to Tamil Named Entity Recognition.

The next phase involves model training using the annotated dataset. Conventional machine learning models like Support Vector Machines coexist with advanced models, emphasizing transfer learning from pre-trained transformer models. Data augmentation techniques enhance the training dataset, promoting better generalization and robustness.

The architecture incorporates comprehensive evaluation metrics, including precision, recall, and F1-score, alongside a multi-level error analysis. This stage is pivotal for understanding the model's performance and identifying areas of improvement. User-provided test data aids in fine-tuning the system based on real-world linguistic variations.

A unique feature of the architecture is its emphasis on continuous development through active learning. Human-in-the-loop feedback and user corrections contribute to iterative model updates. The architecture facilitates adaptability to evolving linguistic nuances and ensures the system's relevance over time.

The project recognizes the challenges posed by limited labelled data for Tamil. The architecture explores resource-efficient training techniques, including semi-

supervised learning, to make optimal use of available resources while maximizing model performance.

Beyond model development, the architecture envisions the creation of an open-source toolkit for Tamil NER. This toolkit includes pre-trained models, data pre-processing scripts, and extensive documentation. The architecture encourages collaboration within the Tamil NLP community, fostering a collective effort towards advancements in the field.

The architecture places a strong emphasis on multi-level error analysis, examining morphological, syntactic, and semantic aspects. This nuanced approach provides valuable insights into the challenges specific to Tamil NER, guiding subsequent improvements and contributing to the project's significance.

The overarching architecture positions this research project at the forefront of linguistic research, machine learning, and community-driven open-source contributions. The outcomes are anticipated to make a lasting impact on NLP practices for the Tamil language, offering a blueprint for future projects aimed at addressing the complexities of morphologically rich languages. The architecture's adaptability and community-centric focus set the stage for continued advancements in Tamil NLP.

METHODOLOGY

The methodology of this research project revolves around the development of a robust Named Entity Recognition (NER) system tailored for the Tamil language. The approach is characterized by a comprehensive methodology that seamlessly integrates traditional machine learning models with state-of-the-art deep learning architectures.

The methodology initiates with the meticulous collection of a diverse dataset in Tamil, representing the linguistic richness of the language. Subsequently, morphological analysis is applied to the raw text data, dissecting words into morphemes to capture the intricate structures inherent in Tamil.

The crux of the methodology lies in the design of a hybrid model that amalgamates Conditional Random Fields (CRF) with transformer-based architectures, such as Bidirectional LSTM-CRF and BERT. This combination aims to leverage the strengths of sequential modelling and contextual understanding for effective entity recognition.

The methodology involves rigorous model training using the annotated dataset. Conventional machine learning models, including Support Vector Machines, undergo training alongside deep learning models. Transfer learning is employed, pre-training transformer models on a large corpus and fine-tuning them on the domain-specific Tamil NER dataset.

Augmenting the dataset is a crucial step to enhance model robustness. The methodology incorporates data augmentation techniques, introducing variations in the training dataset to improve the model's ability to generalize across diverse linguistic patterns.

The evaluation phase employs standard NER metrics such as precision, recall, and F1-score. A distinctive feature of the methodology is the multi-level error analysis, examining errors at morphological, syntactic, and semantic levels. This nuanced approach provides a holistic understanding of model performance.

The methodology adopts an iterative approach to model development through active learning. Human-in-the-loop feedback is sought, enabling continuous refinement of the system. This dynamic feedback loop ensures adaptability to evolving linguistic nuances.

Recognizing the challenges posed by limited labelled data in Tamil, the methodology explores resource-efficient training techniques. Semi-supervised learning is investigated to maximize the utility of available data and optimize model performance.

Beyond model development, the methodology envisions the creation of an open-source toolkit tailored for Tamil NER. This toolkit comprises pre-trained models, data pre-processing scripts, and detailed documentation. The methodology encourages active participation from the Tamil NLP community, fostering collaborative contributions.

The methodology concludes by highlighting the significance of the study. By addressing the complexities of morphologically rich languages, this methodology contributes to the broader understanding of efficient NER systems. Future directions involve the continuous improvement of the NER system, informed by user feedback and evolving linguistic landscapes, emphasizing the project's lasting impact on Tamil NLP practices.

CRF- BERT MODEL

The CRF-BERT model is a hybrid natural language processing (NLP) model that combines Conditional Random Fields (CRF) and BERT (Bidirectional Encoder Representations from Transformers) architectures. This model is developed to address the task of Named Entity Recognition (NER) in the Tamil language.

NER is a fundamental task in NLP that involves identifying and classifying named entities (such as persons, organizations, locations) in a given text. The goal is to recognize and categorize specific entities to extract structured information from unstructured text.

CRF is a type of probabilistic graphical model used for sequence labeling tasks. In the context of NER, it helps model the sequential dependencies between

tokens in a sentence. CRF is particularly useful for capturing contextual information and ensuring that the predicted labels are coherent within a given sequence.

BERT is a powerful pre-trained transformer-based model introduced by Google. It is designed to capture bidirectional contextual information from input text, considering both the left and right context of each word. BERT has demonstrated state-of-the-art performance in various NLP tasks due to its ability to understand the context and relationships between words.

The hybrid nature of the CRF-BERT model involves integrating the strengths of both CRF and BERT. BERT captures the contextual information, while CRF addresses the sequential dependencies. This combination is particularly beneficial for languages like Tamil, where morphological complexities and contextual nuances play a significant role.

The CRF-BERT model is trained on a carefully annotated dataset with labelled entities in Tamil. During training, the model learns to predict the named entity labels for each token in a sequence, taking into account both the local context (captured by BERT embeddings) and the sequential dependencies (captured by CRF).

The model's performance is evaluated using standard NER metrics such as precision, recall, and F1-score. Error analysis is conducted to understand the challenges and areas for improvement. This iterative process helps refine the model and enhance its accuracy.

The CRF-BERT model is a key component of the larger NLP project focused on improving NER for the Tamil language. Its unique architecture, combining CRF and BERT, is tailored to address the linguistic subtleties and complexities of Tamil, resulting in an optimized NER system for the specific requirements outlined in the project abstract.

PROCESS FLOW

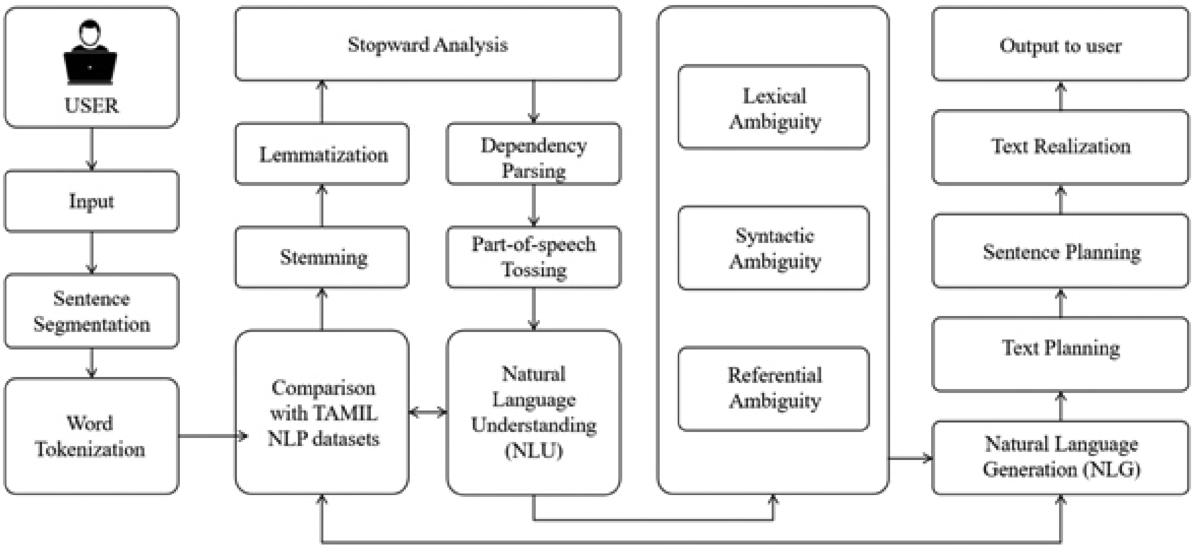


Figure 2: Process flow of the model

Indic NLP Library:

An Indic NLP library would likely focus on providing tools, resources, and algorithms specifically tailored for processing and analyzing text in Indic languages. Indic languages include but are not limited to Hindi, Bengali, Tamil, Telugu, and others.

Indic languages often have rich morphological structures. A dedicated library might include tokenization and morphological analysis components designed to handle the complexities of Indic scripts.

NATURAL LANGUAGE PROCESSING (NLP)

Natural Language Processing (NLP) is a subfield of artificial intelligence that deals with the interaction between computers and humans in natural language. It involves the use of computational techniques to process and analyze natural language data, such as text and speech, with the goal of understanding the meaning behind the language.

User input

User input is crucial for NLP systems to perform tasks such as language understanding, sentiment analysis, information retrieval, and more. The nature of user input can vary widely, and NLP systems need to be capable of handling diverse linguistic expressions and queries.

Sentence segmentation

Sentence segmentation enables NLP systems to understand the structure of the text, making it easier to extract meaning and relationships. It is particularly important for tasks such as machine translation, sentiment analysis, and summarization.

Tokenization

The initial step often involves tokenization, breaking down the user input into smaller units called tokens. Tokens can be words, phrases, or even characters, depending on the level of granularity required for the specific NLP task.

Tamil NLP Datasets

Tamil NLP datasets are collections of text data in the Tamil language that are specifically curated and annotated for various Natural Language Processing (NLP) tasks. These datasets play a crucial role in training and evaluating machine learning models for tasks such as sentiment analysis, named entity recognition, machine translation, and more. They are essential for advancing research and development in the field of Tamil Natural Language Processing.

Lemmatization and Stemming

Lemmatization reduces words to their base or dictionary form, while stemming aims to cut words down to their root form. Both processes help in reducing inflected words to a common base for better analysis.

Dependency parsing

Dependency parsing is the process of analyzing the grammatical structure of a sentence to identify the relationships between words. It involves determining the syntactic dependencies between words and representing them as a tree structure, where each word is a node, and the edges represent the relationships.

POS tagging

Part-of-Speech tagging, also known as POS tagging or grammatical tagging, is the process of assigning grammatical categories (such as noun, verb, adjective, etc.) to each word in a sentence. Each word is labeled with a specific tag indicating its syntactic and grammatical role in the sentence.

NATURAL LANGUAGE UNDERSTANDING (NLU)

Natural Language Understanding is an area of artificial intelligence to process input data provided by the user in natural language say text data or speech data. It is a way that enables interaction between a computer and a human in a way like humans do using natural languages like English, French, Hindi etc.

Lexical Ambiguity

One common source of ambiguity is lexical ambiguity, where a word has multiple meanings. Homonyms and polysemous words can lead to confusion. For example, the word “bank” can refer to a financial institution or the side of a river, and determining the correct interpretation relies on the context in which it is used.

Syntactic Ambiguity

Syntactic ambiguity occurs when the structure of a sentence allows for multiple valid interpretations. For instance, consider the sentence “I saw the man with the telescope.” Here, ambiguity arises as it is unclear whether the speaker used a telescope to see the man or if the man had the telescope.

Semantic Ambiguity

Semantic ambiguity involves multiple interpretations of the meaning of a sentence due to word sense or reference ambiguity. For example, the phrase “She saw the dog on the hill with the telescope” can be interpreted differently based on whether the telescope belongs to the person or the dog.

Pragmatic Ambiguity

Pragmatic ambiguity arises when the intended meaning relies heavily on the context or background knowledge. Anaphoric references and conversational implicatures are common sources of pragmatic ambiguity, where understanding the speaker's intentions requires knowledge beyond the explicit content of the utterance.

NATURAL LANGUAGE GENERATION (NLG)

Natural Language Generation (NLG) is a sub-component of Natural language processing that helps in

generating the output in a natural language based on the input provided by the user. This component responds to the user in the same language in which the input was provided, say the user asks something in Tamil then the system will return the output in Tamil.

Text planning

Text planning is the initial phase in NLG where the system determines the overall structure, content, and organization of the generated text. This involves deciding what information to include, how to order it, and what level of detail is appropriate. Text planning is crucial for creating coherent and contextually relevant messages.

Sentence planning

Sentence planning is the intermediate step in NLG, occurring after text planning. In this phase, the system focuses on generating individual sentences that convey the selected information. It involves deciding how to phrase the information and selecting appropriate sentence structures.

Text realization

Text realization is the final phase in NLG where the system converts the planned and structured information into actual text that can be presented to the user. It involves mapping the abstract representation generated in the previous stages to a grammatically correct and fluent natural language output.

Acknowledgement

I extend my sincere gratitude to Mr. Rajkumar Kalaimani for his invaluable guidance and mentorship throughout this project. Special thanks to my dedicated team members for their unwavering cooperation and collaborative efforts. I also express my appreciation to the Tamil Virtual Academy for providing this incredible opportunity, fostering an environment of learning and innovation. This project has been enriched by the collective contributions of all involved, and I am grateful for the support and encouragement received from these esteemed individuals and organizations.

REFERENCES

[1] David Collins, Alan Deck,Myra McCrickard "Computer Aided Instruction: A Study Of Student Evaluations And Academic Performance", Journal of College Teaching & Learning – November 2008 , Volume 5, Number 11

[2] Bellomo, T. (2009, April)." Morphological analysis and vocabulary development: Critical criteria." Reading Matrix, 9(1),44-55: http://www.readingmatrix.com/articles/bellomo/article.pdf

[3] Joakim Nivre, "Dependency Grammar and Dependency Parsing"

[4] http://stp.lingfil.uu.se/~nivre/docs/05133.pdf

[5] Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P., Rajendran S., "Chunker for Tamil," artcom, pp.436-438, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009, IEEE Press,doi: 10.1109/ARTCom.2009.191

[6] Dhanalakshmi V., Anand Kumar M., Rekha R.U., Arun Kumar C., Soman K.P., Rajendran S., "Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches," artcom, pp.433-435, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009,IEEE Press,doi: 10.1109/ARTCom.2009.184

[7] Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S, "Tamil Part-of-Speech tagger based on SVMTool", In Proceedings of the COLIPS International Conference on natural language processing(IALP), Chiang Mai, Thailand. 2008.

[8] Jes´us Gim´enez and Llu´ıs M`arquez.(2004) "SVMTool: A general pos tagger generator based on support vector machines". In Proceedings of the 4th LREC Conference, 2004.

[9] Lafferty J, McCallum A, Pereira F. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". Proceedings of ICML: 282-289.

[10] Anand Kumar M, Dhanalakshmi V, Soman K.P and Rajendran S. "A Sequence Labeling Approach to Morphological Analyzer for Tamil Language", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 2201-2208.

[11] Anand Kumar M, Dhanalakshmi V, Rekha R U, Soman K.P and Rajendran S. Article: "A Novel Data Driven Algorithm for Tamil Morphological Generator", International Journal of Computer Applications 6(12):52–56, September 2010.

LSTM-based Sequence-to-Sequence Models for Tamil Text Summarization

B Sanjana and Sabari Bala Sundar S

ABSTRACT

Summarizing text from extensive documents presents a challenging task. Various pre-trained models, like IndicBART, are employed to condense lengthy passages. In this research, we aim to construct a Long Short-Term Memory (LSTM) network model based on the seq2seq technique, aiding in Natural Language Processing for Tamil texts. The dataset used in this study comprises Tamil news articles and has undergone preprocessing before applying Word2Vec methods for word embedding. The resulting embeddings are then utilized within the model. Finally, the model's performance is evaluated using specific metrics.

B Sanjana and Sabari Bala Sundar S
Department of Information and Technology
SRM Valliammai Engineering College
Email: sanjanabarath@gmail.com,
sabaribalasundar704@gmail.com

1. INTRODUCTION

The task of effectively condensing extensive documents into concise summaries stands as a significant challenge. The objective of automated headline generation is to succinctly and informatively summarize the text's content. Additionally, it aims to offer potential readers a brief yet comprehensive preview of the material.[1] Conventional methods often grapple with the intricacies and nuances of language, especially when handling diverse and lengthy texts. However, recent strides have ushered in a new era with the advent of sophisticated pre-trained models, vastly enhancing the efficiency and accuracy of text summarization. Despite their prowess, these models encounter limitations, particularly when confronted with underrepresented languages and specific textual formats.

Our proposal centers on a pioneering application of a Long Short-Term Memory (LSTM) network model harnessing the sequence-to-sequence (seq2seq) technique. The crux of our endeavor is to surpass the capabilities of existing pre-trained models when handling the intricate complexities of

Tamil texts. “Automatic Text Summarization (ATS) creates summaries that encompass crucial sentences, ensuring the inclusion of all relevant and vital information from the original document.”[2] PEGASUS [5] is one of the transformer models that is used for abstractive summarization.

By integrating the Word2Vec method for word embedding, our approach revolutionizes the treatment of Tamil news articles, books, and various textual forms. The authors in [6] improved the pointer-generator network to address the Out-of-Vocabulary (OOV) hurdle and accomplished purely abstractive summarization. They incorporated hierarchical attention, utilizing a hierarchical encoder that encompasses both word-level and sentence-level information. This transformation enables the LSTM model to effectively learn and distill information for succinct summarization.

Our methodology encompasses model development, with a meticulous evaluation of the model's performance against specific metrics. Our primary objective is to showcase the potential of a bespoke LSTM model, coupled with Word2Vec embedding, in proficiently

summarizing Tamil documents. This model aims to serve as a compelling alternative to traditional pre-trained models like IndicBART, especially when processing lengthy passages.

In essence, our initiative stands as a testament to the evolving landscape of NLP, propelling Tamil document summarization into a realm where tailored models cater to the intricate nuances of language, offering a more effective and nuanced summarization solution.

2. RELATED WORK

a) Automatic Text Summarization (ATS)

Automatic Text Summarization (ATS) is a sophisticated process involving the extraction of crucial segments from a document, often encapsulating the essence of the entire content. The approach for the Tamil language involves a hybrid model amalgamating Keyword-based scoring, sentiment analysis, and Text Ranking-based scoring to facilitate Automatic Text Summarization in Tamil. The proposed model averaged with an accuracy of around 0.81 as Recall score, 0.61 as Precision score and 0.67 as F score.

b) Extractive Text Summarization (ETS)

The process of text summarization aims in capturing the key information within a document. Abstractive Text Summarization (ATS) and Extractive Text Summarization (ETS) stand as the primary techniques in this field. A Punjabi Extractive Text Summarizer uses an unsupervised machine learning approach that involves distinct modules like Punjabi text tokenization, stop-word elimination, similarity matrix generation, ranking based on this matrix, and the subsequent generation of a concise summary.

c) IndicBart

IndicBART is a specialized tool designed to summarize text passages in Indian languages. It's built on the BART model and is great at understanding languages like Hindi, Bengali, Tamil, and more. This model helps condense information from these languages, making it easier to grasp the main points in texts, especially in Indian languages. IndicBART exhibits superior performance in Hindi text summarization compared to other models like the multilingual T5 variant.

3. PROPOSED SYSTEM

Approach to summarize Tamil articles, utilizing advanced techniques within natural language processing. Our strategy revolves around the adept utilization of an LSTM network operating within a sequence-to-sequence framework—a cutting-edge solution for abstractive summarization.

The methodology kicks off with crucial steps: initial tokenization and embedding. These steps are pivotal, transforming the richness of Tamil text into numerical representations. This transformation lays the groundwork for the subsequent analysis conducted by the LSTM-based encoder-decoder architecture.

The encoder delves into the article's content, extracting its core essence to craft a comprehensive context vector. This vector encapsulates the fundamental concepts and nuances, providing a foundation for the summarization process.

The decoder is equipped not only with the context vector but also supported by an attention mechanism—a pivotal intelligence booster. With this, the decoder intelligently sculpts a concise summary, adeptly highlighting crucial segments of the original article. This attention-driven approach ensures that the summarization focuses on the most pertinent elements, maintaining the integrity and relevance of the content.

But mastery doesn't come without effort. Our model undergoes rigorous iterative training and optimization using paired data. Through this iterative process, the model evolves, honing its ability to produce coherent, informative summaries that retain the linguistic richness inherent in the Tamil language.

In essence, our approach amalgamates sophisticated technology and linguistic finesse to create a model that adeptly distills the essence of Tamil articles. This model not only condenses the information but does so intelligently, ensuring the summary reflects the depth and integrity of the original content.

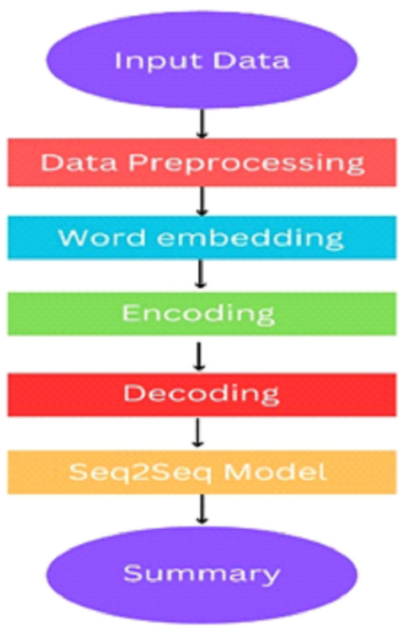


Fig. 1: Sequence Diagram

4. METHODOLOGY

a) Data Collection

The dataset originates from Kaggle, a popular data Science platform, featuring five columns housing diverse information. Among the attributes within these columns are: 'news_id,' serving as a unique identifier; 'news_date,' capturing the temporal aspect of the data; 'news_category,' delineating the thematic classification; and 'news_title' along with 'article,' providing textual insights into the news content. This dataset presents a rich reservoir of information, ideal for exploratory analysis and potentially fostering insights into various facets of news-related domains.

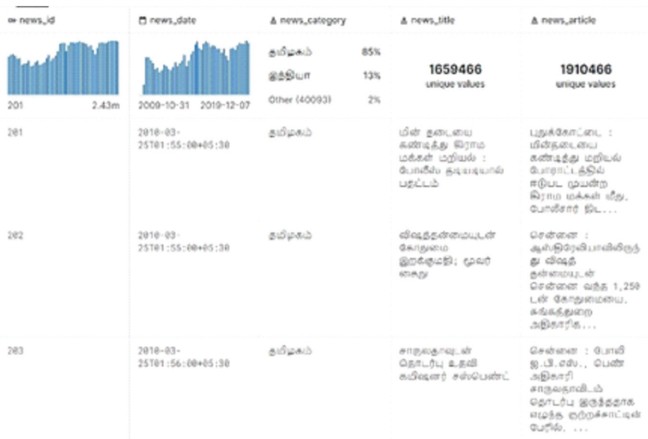


Fig. 2: Dataset

b) Data Preprocessing

To prepare the data for model utilization, initial processing is essential. To accomplish this task, a data processing pipeline has been developed. The inaugural stage of this pipeline focuses on data cleaning. Certain attributes are unnecessary for observation and analysis. It is best to construct a model with cleaned data. The attributes that are unnecessary for the processing are removed. The unwanted columns are new_id, news_date and news_category. The fields that are empty are also removed to ensure that it doesn't impact the final result. As part of the abstract approach, we won't be removing or omitting stop words that might cause loss of information.

c) Word embedding

Word embedding is a technique in natural language processing that converts words into numerical vectors. These vectors capture semantic relationships, representing words' meanings and contexts in a multi-dimensional space. Gensim is a popular Python library that facilitates the creation of Word2Vec models by transforming words into high-dimensional vectors, preserving their semantic relationships. Through continuous training, the model captures intricate word associations and similarities, enabling it to map semantic meanings in a multi-dimensional space.

d) Seq2Seq Model

The seq2seq model, utilizing Long Short-Term Memory (LSTM) networks, is a fundamental architecture for summarizing Tamil articles. This technique comprises two essential components: an encoder and a decoder. The LSTM-based encoder processes the input Tamil article, converting it into a fixed-size context vector while capturing its essential information. This context vector encapsulates the semantic essence of the article. Subsequently, the LSTM-based decoder takes this context vector as input and generates a concise summary sequentially. The decoder attends to different parts of the encoded article through an attention mechanism, allowing it to focus on relevant segments when producing each word of the summary.

e) Encoding

In the summarization of Tamil articles with an LSTM-based seq2seq model, the encoding process involves converting the input text into numerical representations. This begins with tokenization, breaking down the Tamil article into smaller units, such as words or sub words, which are then embedded into high-dimensional vectors to capture their semantic meanings. The LSTM-based encoder then processes these embedded representations, sequentially analyzing the input text to create a context vector. This context vector encapsulates the essential information of the article, forming a foundation for the decoder to generate a summary. Through this encoding phase, the LSTM network learns to distill the salient details of the Tamil article into a condensed numerical format, enabling the subsequent decoding step to craft a coherent and concise summary.

f) Decoding

The decoding process involves using the context vector generated by the encoder as a foundation to generate a concise summary. The LSTM-based decoder, initialized with the context vector, begins the sequence generation for the summary. Employing an attention mechanism, the decoder attends to different parts of the encoded input text while generating each word of the summary. It predicts and generates words one by one, considering the context vector and the learned relationships within the input article.

5. RESULTS

Deep learning models demand extensive data for robust performance, especially within the encoder-decoder architecture used for text summarization. To harness their true potential, a substantial dataset was curated, predominantly comprising online Bangla

news articles. The model required sequences of text for training: utilizing full news articles as input and using their titles as reference summaries for abstractive summarization.

Training an abstractive summarization model in natural language processing poses challenges as the machine aims to generate summaries not explicitly present in the original text. To accomplish this, probability calculations play a critical role, shaping the machine's output based on maximum likelihood. Preprocessing the input data and training it on a deep learning model, specifically leveraging Tensor Flow 2.0.2, were crucial steps in this experiment.

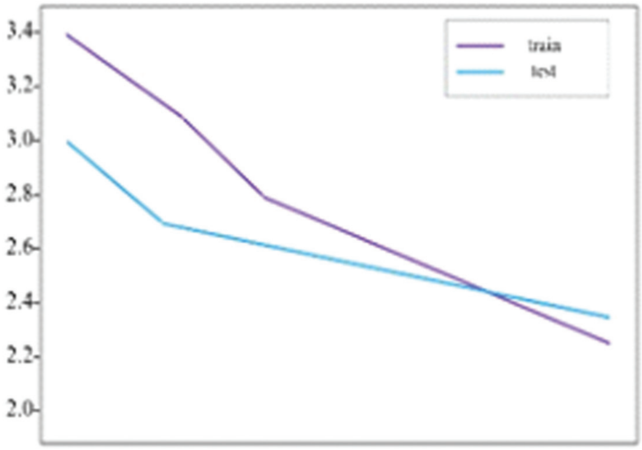


Fig. 3: Result

The training parameters were meticulously fine-tuned, considering factors like cluster size, learning rate, and the number of layers, all of which significantly impact the model's performance.

Minimizing loss during training was prioritized, employing the "RMSProp" optimizer within the model architecture. The utilization of high-end GPUs, particularly through Google Colab's GPU services, significantly expedited the training process.

The experiment utilized a latent dimension of 300, embedding dimension of 200, batch size of 345, 50 hidden units, with 27 epochs (with early stops). The performance evaluation using ROUGE metrics indicated promising results: ROUGE-1 at 0.60, ROUGE-2 at 0.47, and ROUGE-L at 0.45. These metrics signify enhanced similarity between system-generated summaries and reference summaries in the dataset, showcasing the model's proficiency in generating accurate and coherent summaries. However, it's essential to note the limitations of automatic evaluation metrics, particularly their inability to fully capture nuances of meaning. Nonetheless, for an abstractive approach, the model's ability to generate comprehensive and informative new sentences remains a notable achievement. The authors in [3] introduced LongT5, exploring the impacts of scaling context lengths and model sizes to attain cutting-edge outcomes. Additionally, they conducted meticulous human annotations, as outlined in [4], to assess abstractive summarization models, striving to enhance the ROUGE score.

6. CONCLUSION

The paper signifies a remarkable advancement within the realm of Natural Language Processing (NLP), particularly in tackling the intricate task of summarizing documents written in the Tamil language. Despite the strides made by traditional NLP methodologies and pre-trained models, their limitations become evident when confronted with the intricate complexities and idiosyncrasies inherent in less-represented languages such as Tamil. The crux of our paper lies in the introduction of a novel model—one that merges a Long Short-Term Memory (LSTM) network with the sequence-to-sequence (seq2seq) technique, fortified by the implementation of Word2Vec for word embedding. This tailored approach is specifically engineered to confront and overcome the challenges posed by the unique linguistic landscape of Tamil. Traditional approaches often falter in capturing the essence and subtleties of less-represented languages. Our model, however, represents a paradigm shift by addressing these limitations head-on. By harnessing the capabilities of LSTM networks and seq2seq architecture, combined with the efficiency of Word2Vec embedding, our model empowers itself to navigate through the intricate nuances, varied sentence structures, and contextual complexities intrinsic to Tamil.

In essence, this proposed model stands as a beacon of innovation within the field of NLP, offering a tailored solution that transcends the limitations of traditional approaches. It signifies a pivotal step towards enabling more effective and nuanced document summarization specifically for languages like Tamil, thus contributing significantly to the evolution of NLP technology in handling underrepresented linguistic domains.

7. FUTURE ENHANCEMENT

The incorporation of advanced deep learning methodologies, particularly Transformer models, represents a significant leap forward in enhancing the accuracy and contextual understanding within text summarization models. The intrinsic architecture of Transformer models, notably their attention mechanisms, empowers these systems to capture intricate relationships and dependencies within textual data more effectively.

In the realm of Tamil text summarization, leveraging Transformer models can be transformative. Training these models with diverse datasets in Tamil plays a pivotal role in honing their performance. The exposure to a wide array of textual sources enables these models to grasp the nuances, variations, and complexities inherent in the language.

By exposing Transformer models to diverse data, they become adept at understanding the intricate context and subtleties within Tamil text. This exposure enriches their understanding of sentence structures, semantics, and contextual cues, ultimately leading to heightened accuracy and proficiency in generating coherent and informative summaries.

In essence, the integration of Transformer models, coupled with training on diverse Tamil datasets, marks a significant stride in the evolution of text summarization. This combination not only elevates the accuracy levels but also enhances the model's ability to grasp the essence and context of Tamil text, paving the way for more refined and nuanced summarization outputs.

REFERENCES

[1] G. Sharma and D. Sharma, "Automatic Text Summarization Methods: A Comprehensive Review," SN Comput Sci, vol. 4, no. 1, pp. 1–18, Jan. 2023, doi: 10.1007/S42979-022-01446-W/METRICS.

[2] A. P. Widyassari et al., "Review of automatic text summarization techniques & methods," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 4, p.

[3] Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y. H., & Yang, Y. (2021). LongT5: Efficient text-to-text transformer for long sequences. arXiv preprint arXiv:2112.07916.

[4] Koh, H. Y., Ju, J., Zhang, H., Liu, M., & Pan, S. (2022). How Far are We from Robust Long Abstractive Summarization? arXiv preprint arXiv:2210.16732.

[5] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp.11328-11339). PMLR.

[6] N. Dhar, G. Saha, P. Bhattacharjee, A. Mallick, and M. S.Islam, "Pointer over attention: An improved bangla text summarization approach using hybrid pointer generator network," CoRR, vol. abs/2111.10269, 2021.

[7] S. Abujar, A. K. M. Masum, M. S. Islam, F. Faisal, and S. A. Hossain, "A bengali text generation approach in context of abstractive text summarization using rnn," 2020.

[8] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., ... & Sifre, L. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

[9] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

[10] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

[11] Wang, B., & Komatsuzaki, A. (2022). GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model, 2021.

[12] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, "Neural abstractive text summarization with sequenceto-sequence models," ACM/IMS Transactions on Data Science, vol. 2, pp. 1–37, 2 2021.

[13] H. Zhang, H. Song, S. Li, M. Zhou, and D. Song, "A survey of controllable text generation using transformer-based pre-trained language models," ArXiv, vol. abs/2201.05337, 2022.

[14] O. Sen, M. Fuad, M. N. Islam, J. Rabbi, M. Masud, M. K. Hasan, M. A. Awal, A. A. Fime, M. T. H. Fuad, D. Sikder, and M. A. R. Iftee, "Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods," IEEE Access, vol. 10, pp. 38999–39044, 2022.

[15] S. Abujar, A. K. M. Masum, M. S. Islam, F. Faisal, and S. A. Hossain, A Bengali Text Generation Approach in Context of Abstractive Text Summarization Using RNN, pp. 509–518. 2020.

[16] K. Prudhvi, A. B. Chowdary, P. S. R. Reddy, and P. L. Prasanna, Text Summarization Using Natural Language Processing, pp. 535–547. 2021.

கணினி வழி இலக்கிய மொழி ஆய்வு

முனைவர் ப. டேவிட் பிரபாகர்

ஆய்வுச்சுருக்கம்

தமிழில் செம்மொழிப் பனுவல்கள் உள்ளிட்ட இலக்கியத் தரவகங்கள் உருவாக்கப்பட்டுள்ளன. இலக்கியம் ஒரு மொழிக் கலை என்னும் அடிப்படையில் இயற்கை மொழியாய்வுக் கருவிகளின் வழி இலக்கியப் பனுவல்களை விரிவாக ஆராய இயலும். இலக்கணக் குறியீடு பெற்ற இலக்கியத் தரவகங்கள், சொல்நிலையில் தமிழில் உருவாக்கப்பட்டுள்ளன. ஆயினும், சொற்சேர்க்கை, தொடரியல் – பொருண்மையியல் உறவுகள், சூழ்பொருள் ஆகிய நிலைகளில் இத்தகைய தரவகங்கள் இன்னும் மேம்படுத்தப்பட வேண்டிய நிலையில் உள்ளன. அதே வேளையில், இலக்கியப் பனுவல்களை மேம்பட்ட, ஒருங்கிணைந்த கணினியப் பார்வையில் அணுகி ஆய்வதற்காகக் களங்களை இவ்வாய்வுரை முன்மொழிகிறது.

தொடை நயம் உள்ளிட்ட யாப்பியல் கூறுகள், சந்தப் பாடல்களின் இசைமை ஒழுங்கு – பொருண்மை உறவு போன்றவற்றைச் சொல்லியல் நிலையில் கண்டறிய இயலும். இதுபோன்றே, தொடர் நிலையில், தொடர் அமைப்புக் கட்டுப்பாடுகள், விலகல்கள், மீறல்கள் முதலியவற்றின் அடிப்படையில் அமையும் இலக்கிய உத்திகளைக் கண்டறிய முடியும்.

உள்ளடங்கு உறவு (Hyponymy), உள்ளடக்கு உறவு (Hypernym) சினை–முதல் உறவு (Part–whole relation), உட்படுத்து உறவு ஆகியவற்றை வெளிப்படுத்தும் சொல்வலை (wordnet) வாயிலாகக் கருப்பொருள் – உரிப்பொருள் – திணை உறவுகள், சொல் வடிவம் – கூற்று – துறை – படைப்பாளி உறவு முதலியவற்றைக் கண்டறிய இயலும். மேம்பட்ட நிலையில் உவமை, உருவகம், முரண் முதலிய அணிகள், தொடரிணைவு (cohesion) கருத்திணைவு (coherence) முதலிய எடுத்துரைப்பியல் உத்திகள், பேச்சுச் செயல்பாடுகள், நாடகப் பாங்கு ஆகியவற்றையும் கணினி வழி கண்டறிய இயலும்.

எனவே இயற்கை மொழி ஆய்வு மற்றும் விரிதரவுக்கான கருவிகளை இலக்கியத் தரவகங்களில் மேம்பட்ட நிலையில் கையாள்வதன் வழி இலக்கியப் பனுவல்களின் கலை நுட்பங்களை நன்கு வெளிப்படுத்த இயலும்.

முனைவர் ப. டேவிட் பிரபாகர்
தமிழ்த்துறைத் தலைவர், சென்னைக் கிறித்தவக் கல்லூரி
மின்னஞ்சல்: tamilprofessor@gmail.com

அறிமுகம்

கணினி வழி இலக்கிய ஆய்வு இலக்க மாந்தவியலின் (Digital Humanities) துணைப் புலமாக விளங்குகிறது. இலக்கியப் பனுவல்களைக் கணினி வழி பகுப்பாய்வதன் வழி இலக்கியங்களின் வடிவம், நடை, பொருண்மை, அழகியல் கூறுகள், இவற்றுக்கு இடையே நிலவும் உறவுகள் முதலியவற்றை வெளிப்படுத்த இயலும். அணுக்க அல்லது ஆழ்ந்த வாசிப்பின் வழி மேற்கொள்ளப்படும் மரபார்ந்த இலக்கிய ஆய்வுக்கு மாற்றாகக் கணினி வழி இலக்கிய ஆய்வு முறை அமையும்.

இயற்கை மொழி ஆய்வில் பல்வகை மொழிக் கருவிகள் பயன்படுத்தப்படுகின்றன. சொல்லியல், தொடரியல் நிலைகளில் பனுவல்கள் பகுப்பாய்வுக்கு உட்படுத்தப்படுகின்றன. இதன் வழி இலக்கியப் பனுவல்களின் நடையியல் கூறுகளை இனம் கண்டு விவரிக்க இயலும். இலக்கியப் பனுவல்களில் ஒலி அமைப்பு, சொல் அமைப்பு, தொடர் அமைப்பு ஆகியன ஒருங்கிணைந்து ஓர் உயிரியாய் விளங்குவதால், இலக்கியப் பனுவல்களைக் கருத்தாடல் நிலையில் (Discourse Approach) அணுகும்பொழுது, திறன்மிகுந்த நிலையில் இலக்கியத் திறனாய்வுக்குக் கணினியைப் பயன்கொள்ள இயலும்.

சங்க இலக்கியம் தொடங்கி நவீனப் புனைகதைகள் வரையில் பல்வேறு இலக்கிய விரிதரவுகள் தமிழில் உருவாக்கப்பட்டுள்ளன. தமிழில் இலக்கணக் குறியீடு செய்யப்பட்ட இலக்கிய உரைத் தரவுகள் உருவாக்கப்பட்டுள்ளன. தமிழ் இணையக் கல்விக்கழகம் தமிழ் இலக்கியங்களின் விரிதரவைச் சொல்லியல், தொடரியல், பொருண்மையில் விளக்கங்களோடு உருவாக்குவதில் முன்னோடிப் பணிகளை மேற்கொண்டுள்ளது; மொழி ஆய்வுக்கான கருவிகள் சிலவற்றையும் உருவாக்கியுள்ளது. செம்மொழி மத்திய தமிழாய்வு நிறுவனமும் உ.வே.சா. செம்மொழித் தரவகத்தை உருவாக்கி வெளியிட்டுள்ளது. தொடரியல், பொருண்மை உறவுகள், சூழ் பொருள் சார்ந்து இத்தகைய தரவகங்கள் மேம்படுத்தப்பட வேண்டிய நிலையில் உள்ளன. மேலும் இலக்கிய உரை தரவுகளின் புறக்கூறுகளாகக் கருதத்தக்க திணை, கூற்று, துறை, புலவர், நூல், முதலியவற்றை ஒருங்கிணைத்து இலக்கிய ஆய்வை மேற்கொள்வது நுட்பமான முடிவுகளைப் பெறத் துணைசெய்யும்.

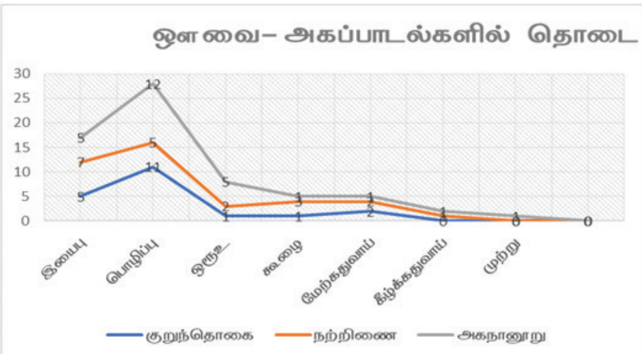
கருவையும் உருவையும் அமைப்பாக்கம் செய்வன இலக்கியங்கள். ஒலிக்கோலம், சொல் தேர்வு, தொடர்க் கட்டு, தொடர்பாடல் அலகு, ஒருங்கிணைவு, கருத்திணைவு, உள்நோக்குடைமை, ஏற்பு, சூழல் இயைபு, பிற பிரதி தொடர்பு முதலியன இலக்கிய மொழியியல் பார்வையில் இன்றியமையாதவை. செய்யுளுக்கு உறுப்பாக அமையும் 34 கூறுகளைத் தொல்காப்பியர் ஒருங்கிணைந்த பார்வையில் அணுகுவதைச் செய்யுளியல் விவரிக்கிறது.

இயற்கை மொழியாய்வுக் கருவிகள் மற்றும் விரிதரவு மொழியாய்வுக்கென உருவாக்கப்பட்டுள்ள கருவிகளின் துணையுடன் இலக்கியப் பனுவல் ஆய்வை முன்னெடுக்க இயலும். சங்கப் பாடல்களின் விரிதரவை அணுகும் கணினி மொழி ஆய்வு முறைகளை இக்கட்டுரை முன்மொழிகிறது.

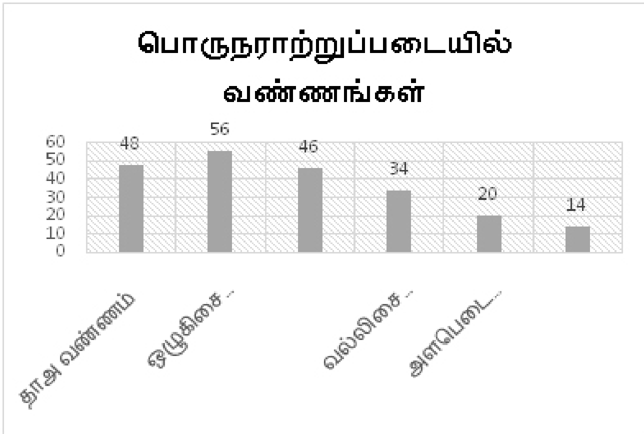
ஒலிய நிலை இலக்கிய ஆய்வு

தொல்காப்பியச் செய்யுளியல் குறிப்பிடும் மாத்திரை, எழுத்து, அசை, சீர், தூக்கு, தொடை, வண்ணம் முதலியன ஒலியக் கூறுகளாகக் கொள்ளத்தக்கவை. இவை செய்யுளின் ஓசை நயத்திற்குக் காரணமாக அமைவதோடு, செய்யுளின் பொருண்மையோடும் தொடர்புடையவையாக அமைகின்றன.

எதுகை, மோனை, இயைபு, அளபெடை ஆகியன ஒலி நிலையிலும் முரண் சொல் நிலையிலும் ஒரூஉ, பொழிப்பு ஆகியன தொடர் நிலையிலும் செய்யுளுக்கு அணியாக அமைகின்றன. பின்வரும் விளக்கப்படம் ஔவையின் அகப்பாடல்களில் காணலாகும் தொடைகளின் வருகையைக் காட்டுகிறது.



ஒரே ஒலி அல்லது அதற்கு இனமான ஒலி, ஓர் அடி அல்லது பாடல் முழுவதும் பயின்று வருவது வண்ணம் எனும் உத்தியாகும். 'வண்ணம் என்பது பாவின்கண் நிகழும் ஓசை விகற்பம்' எனத் தொல்காப்பிய உரையில் பேராசிரியர் குறிப்பிடுகிறார் (1975:116). வண்ணம் என்பது உணர்ச்சியுடன் கூடிய பொருளுடன் தொடர்புடையது. ஆனால் ஓசை என்பது எழுத்து அளவை முதலியன காரணமாக இசையாப்பில் ஏற்படுகிற இசை வேறுபாடு என மேற்கோள் காட்டும் பா.வீரப்பன் (1989:52), பத்துப்பாட்டில் இடம்பெற்றுள்ள வண்ணங்களை எடுத்துக்காட்டியுள்ளார். பொருளுக்கு ஏற்ப வண்ணங்கள் பயன்கொள்ளப்பட்டுள்ள முறையையும் அவர் விவரித்துள்ளார் சிறுபாணாற்றுப்படையில் விறலியரின் களைத்து ஓய்ந்த நடையில் அளபெடை வண்ணம் அழகு செய்வதையும், பெரும்பாணாற்றுப்படையில் முதுவேனிலை விளக்க வல்லிசை வண்ணமும் யாழின் இனிய ஓசையை விவரிக்க மெல்லிசை வண்ணமும் ஆளப்பட்டிருப்பதை பா.வீரப்பன் சுட்டுகிறார் (1989:55). பின்வரும் விளக்கப்படம் பொருநராற்றுப்படையில் மிகுதியாகக் கையாளப்பட்டுள்ள வண்ணங்களின் வருகையைக் காட்டுகிறது.



பாடல்களில் காணலாகும் பல்வேறு ஒலிக்கோலங்கள் பாடலின் உருவையும் கருவையும் தீர்மானிப்பதில் இன்றியமையாத பங்குவகிக்கின்றன. பல்வகைத் தொடை நயங்கள், வண்ணங்கள், ஒலிக்குறிப்புச் சொற்கள் (அச்சம், விரைவு, செறிவு) ஆகியவற்றைப் பாடலின் பொருண்மையோடு தொடர்புபடுத்திக் கணினி வழி விளக்கலாம்.

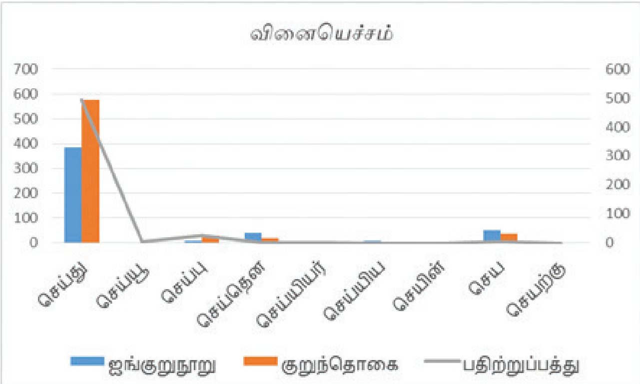
சொல்நிலை இலக்கிய ஆய்வு

சங்கப் பனுவல்களுக்கு இலக்கணக் குறியீடு கொண்ட விரிதரவுகள் கிடைக்கின்றன. சொல் வகைகள் ஆய்வின் தேவைக்கேற்ப வரையறுத்துக் கொள்ளப்படலாம். சொற்களின் வருகை, அவற்றின் பகிர்வு, சொல் வகைகளுக்கிடையே நிலவும் உறவு, சொல் வகைகள் இணைந்து உருவாக்கும் வாய்பாடுகள் (Patterns) போன்றவற்றைக் கண்டறிவது இலக்கிய ஆய்வில் இன்றியமையாதது.

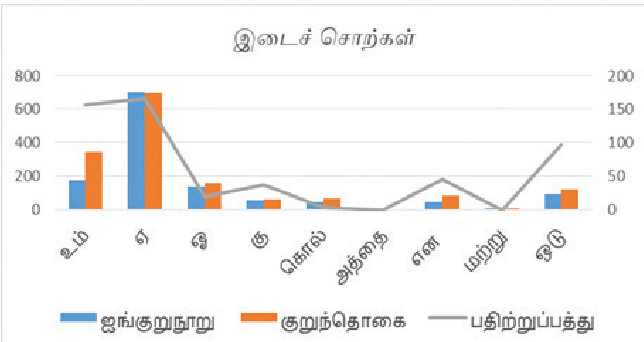
படைப்பின் மொழி ஆளுமை தன்னியலாக வெளிப்படுவது. இது காலம், பாடுபொருள், புலவர், திணை, கூற்று, துறை ஆகியவற்றின் அடிப்படையில் வெவ்வேறு அளவில் மாறுபடுவதற்கு வாய்ப்புள்ளது.

சங்கப் பனுவல்களில் இடம் பெற்றுள்ள சொற்களின் வருகையையும் பகிர்வையும் நூல், புலவர், திணை, கூற்று போன்றவற்றோடு ஒப்பிட்டுநோக்கி கணினி வழி ஆராய இயலும்.

எடுத்துக்காட்டாக, பதிலிப் பெயர்கள் அறத்தொடு நிற்றல் துறையில் அதிகம் ஆளப்பட்டுள்ளதையும். இத்துறைப் பாடல்களில் விளியின் பயன்பாடும் கூடுதலாக அமைந்துள்ளதையும். பெயரடைகளின் பயன்பாடு அறத்தொடுநிற்றல் துறையில் குறைவாக இடம்பெறுவதையும். திரிபுச் சொற்களை ஆளும் முறையில் அறத்தொடுநிற்றல் துறையிலும் வரைவிடைப் பிரிவு துறையிலும் ஒத்த பான்மை வெளிப்படுவதையும். விளியைப் போன்றே அறத்தொடுநிற்றல் துறையில் வியங்கோளின் வருகையும் மிகுந்துள்ளதையும் கட்டுரையாளரின் ஆய்வு வெளிப்படுத்துகிறது. (ப.டேவிட் பிரபாகர்: 2018). பின்வரும் விளக்கப்படம் ஐங்குறுநூறு, குறுந்தொகை, பதிற்றுப்பத்து ஆகிய நூல்களில் ஆளப்பட்டுள்ள வினையெச்சங்களின் வருகையைக் காட்டுகிறது.



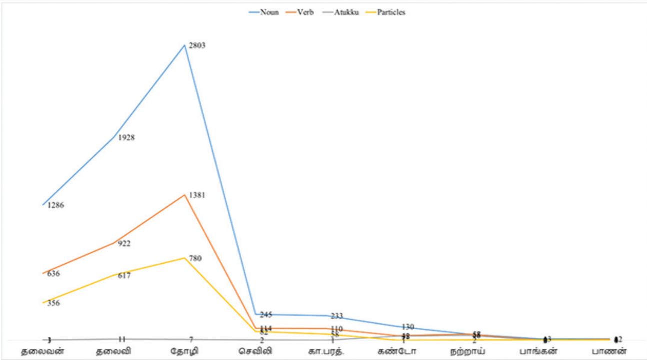
பின்வரும் விளக்கப்படம் ஐங்குறுநூறு, குறுந்தொகை, பதிற்றுப்பத்து ஆகிய நூல்களில் ஆளப்பட்டுள்ள இடைச்சொற்களின் வருகையைக் காட்டுகிறது.



தனிச் சொல் வகைகளின் வருகையின் அடிப்படையில் மட்டுமின்றி, ஒவ்வொரு சொல் வகையின் அலகேற்புப் பண்பின் அடிப்படையிலும், சங்கப் பனுவல்களின் மொழி நடைப் பண்புகளைக் கண்டறியவியலும். அதாவது, ஒவ்வொரு சொல் வகையும் குறிப்பிட்ட சொல் வகைகளையே தமக்கு முன்னும், பின்னும் இடம்பெற அனுமதிக்கின்றன.

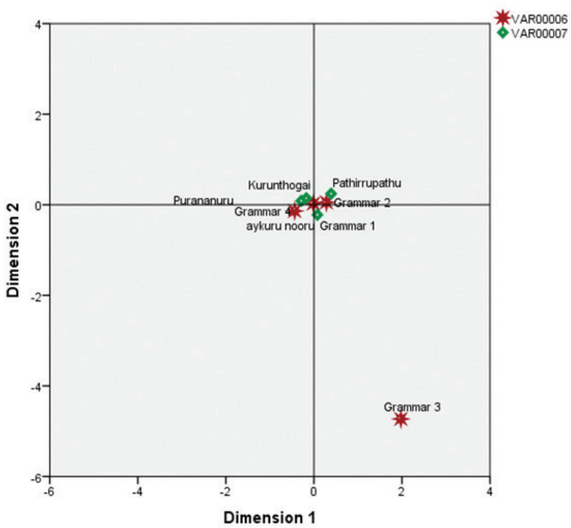
சங்கப் பனுவல்களில் இடம் பெறும் பல்வேறு சொல்வகைகள், அவற்றின் நிகழ்வெண் ஆகியவற்றைக் கண்டறிவதோடு அவற்றின் சராசரி, திட்டவிலக்கம் போன்றவற்றை ஆராய்வதோடு, பனுவல் பகுதிகளை ஒவ்வொரு கொத்தாகக் கொண்டு பிற பகுதிகளுடன் ஒப்பிடும் தொகுதிப் (கிளஸ்டர்) பகுப்பாய்வைப் புள்ளியியல் பகுப்பாய்வுக் கருவிகளின் அடிப்படையில் மேற்கொள்ளலாம்.

பெயர், வினை, இடை/உரி, அடுக்குத் தொடர்/ஒலிக்குறிப்பு எனச் சொல் வகைகளை நான்காகப் பகுத்துக்கொண்டு அவை அகமாந்தர் கூற்றுகளில் இடம்பெற்றுள்ள முறையைப் பின்வரும் படம் காட்டுகிறது.



குழும விவரிப்பு மேற்கொண்டதில், ஆய்வுக்கு எடுத்துக்கொண்ட நான்கு நூல்களில் குறுந்தொகையும், புறநானூறும் அணுக்கமான உறவில் அமைந்திருப்பதைக் காணமுடிகிறது.

பதிற்றுப்பத்தும் ஐங்குறுநூறும் கீழ் மேல் அடுக்குகளில் அமைந்து உறவை வெளிப்படுத்துகின்றன. சொல் வகைகளைப் பயன்படுத்தும் முறையில் பெயர், வினைத் தொகுதிச் சொற்களை ஆளும் முறையில் அமையும் உறவு அணுக்கமாகவும் இடைச் சொற்களை ஆளும் முறையிலான உறவு சற்று விலகியும், அடுக்குத்தொடர்/ஒலிக்குறிப்புச் சொற்களை ஆளும் முறையிலான உறவு மிகவும் விலகியும் அமைந்திருப்பதைக் காணமுடிகிறது.



Word2vec என்பது சொற்களின் ஆளுமைப் போக்கைக் கண்டறியும் இயற்கை மொழி ஆய்வுக்கான (NLP) ஒரு நுட்பமாகும். இது ஒவ்வொரு தனித்த

சொற்களின் திசை வழியை எண்களின் பட்டியலாகத் தருகிறது. இதன்வழி, விரிதரவில் சொற்களின் பொருள் மற்றும் சூழல் சார்ந்த பயன்பாடு பற்றிய தகவல்களைப் பெறலாம். மேலும், சொற்களின் தொடரியல் மற்றும் பொருண்மையியல் பண்புகள், உறவுகளையும் இது காட்டுகிறது.

தொடரன் நிலை இலக்கிய ஆய்வு

எச்சத்தொடர், அடைத் தொடர், விளித்தொடர், வினாத்தொடர், ஏவல் தொடர், உவமைத் தொடர், முரண் தொடர், எதிர்மறைத் தொடர், திருப்புரை எனப் பல்வேறு முறைகளில் இலக்கியப் பனுவல்களில் காணலாகும் தொடர்களைக் கணினி வழி கண்டறிந்து அவற்றைப் பாடலின் கரு, உரு ஆகியவற்றுடன் தொடர்புபடுத்தி விளக்க முயலலாம். வாக்கியங்களின் நீளம், வாக்கியத்தின் அமைப்பு (பொருள்கோள், அடைவிரிப்பு, ஒப்புரை, உருவக நிலை) ஆகிய நிலைகளிலும் மொழி ஆளப்பட்டுள்ள முறையைக் கண்டறிவதற்கான விதிமுறையாக்கங்களை மேற்கொள்ளலாம்.

கருத்தாடல் நிலை இலக்கிய ஆய்வு

கருத்தாடல் நிலையில் இலக்கியப் பனுவல்களை ஆராய்வது சவால் நிறைந்த பணியாகும். ஒலி, சொல், தொடர் ஆகிய நிலைகளைக் கடந்து உரைகளுக்கிடையே (Utterance) அமையும் கருத்திணைவு (Coherence) தொடரிணைவு (Cohesion) ஆகியவற்றைக் கருத்திற்கொண்டு இலக்கியப் பனுவலை முழுமையாகப் பார்க்கக் கருத்தாடல் அணுகுமுறை வழிவகுக்கிறது.

தலைப்பு மாதிரி ஆக்கம் (Topic Modelling) விரிதரவின் உட்பொதிந்துள்ள அல்லது மறைந்துள்ள பொருண்மையியல் கூறுகளை வெளிக் கொணர்வதாகும். இதற்கு உள்ளுறை பொருண்மையியல் ஆய்வு (Latent semantic analysis) எனும் இயற்கை மொழியாய்வு உத்தி துணை செய்யக்கூடியது. ஓசை அடிப்படையிலும், சொல் நிலையிலும் பொருண்மை அடிப்படையிலும் ஒரு பாடலில் அமையும் மொழிக் கூறுகளை இணைத்து நோக்குவதின் வழி பயனுள்ள இலக்கிய ஆய்வைக் கணினி வழி மேற்கொள்ள இயலும். இலக்கியப் பனுவலில் வெளிப்படும் உணர்வு மற்றும் உணர்ச்சி சார்ந்த ஆய்வையும் (sentiment and emotional analysis) முயலலாம்.

சொல்வலை (Wordnet) நுட்பங்களைப் பயன்படுத்திப் பாடல்களில் இடம்பெறும் இணைச்சொல், எதிர்ச்சொல், உள்ளடங்குச் சொல், உள்ளடக்குச் சொல், சினை–முதல் உறவுச்சொல் ஆகியவற்றைச் சொல் நிலையிலும் பொருண்மை நிலையிலும் ஆராயவியலும்.

செய்ந்நன்றி அறிதலும்1, சிற்றினம் இன்மையும்1,
இன் முகம் உடைமையும்1, இனியன் ஆதலும்1,
செறிந்து விளங்கு சிறப்பின் அறிந்தோர்2 ஏத்த;
அஞ்சினர்க்கு அளித்தலும்1, வெஞ் சினம் இன்மையும்1,
ஆண் அணி புகுதலும்1, அழிபடை தாங்கலும்1,
வாள் மீக் கூற்றத்து வயவர்2 ஏத்த;
கருதியது முடித்தலும்1, காமுறப் படுதலும்1,
ஒரு வழிப் படாமையும்1, ஓடியது உணர்தலும்1,
அரி ஏர் உண்கண் அரிவையர்2 ஏத்த;
அறிவு மடம் படுதலும்1, அறிவு நன்கு உடைமையும்1,
வரிசை அறிதலும்1, வரையாது கொடுத்தலும்1,
பரிசில் வாழ்க்கைப் பரிசிலர்2 ஏத்த

மேற்காணும் சிறுபாணாற்றுப்படை பாடல் அடிகளில் இடம்பெற்றுள்ள சொற்களின் ஆளுமை கவிதை நயத்துக்குக் காரணமாக அமைகிறது.

அறிதலும், இன்மையும், உடைமையும், ஆதலும், அளித்தலும், இன்மையும், புகுதலும், தாங்கலும், முடித்தலும், காமுறப் படுதலும், படாமையும், உணர்தலும் ஆகிய சொற்கள்(1) பாடலின் ஓசை மற்றும் பொருண்மையைக் கட்டமைப்பதைக் காண முடிகிறது. இதுபோன்றே, உரிய இடைவெளியில் ஆளப்பட்டுள்ள அறிந்தோர் ஏத்த, வயவர் ஏத்த, அரிவையர் ஏத்த, பரிசிலர் ஏத்த ஆகிய தொடர்களும்(2) பாடலுக்கு அணி சேர்க்கின்றன.

பாடினியின் கேசாதிபாத வருணனையாக அமைந்துள்ள பின்வரும் பொருநராற்றுப்படையின் பாடல் அடிகளில் இடம்பெற்றுள்ள உடல் உறுப்புப் பெயர்களின் ஆளுமை பாடலுக்குப் பொருண்மை சார்ந்த வடிவத்தை வழங்குகிறது.

அறல் போல் கூந்தல்1, பிறை போல் திரு நுதல்1,
கொலை வில் புருவத்து1, கொழுங் கடை
மழைக் கண்1,
இலவு இதழ்1 புரையும் இன் மொழித் துவர் வாய்1,
பல உறு முத்தின் பழி தீர் வெண் பல்1
(பொருநராற்றுப்படை)

கணினி வழி இலக்கிய ஆய்வைப் புனைகதைப் பனுவல்களுக்கும் விரிவாக்கம் செய்ய இயலும். கதைக்கூறு, இழை பொருள், கதைப்பின்னல் ஆகியவற்றுக்கு உருவவியலில் அணுகுமுறையில் சிறப்பிடம் உண்டு. நிகழ்வுகள் பொருண்மையியலின் ஆக்கக் கூறுகளாக விளங்குபவை. நிகழ்வுகளின் வருகை, நிகழ்வுகளின் கோவை, எடுத்துரைப்பியல் முறைமை, பேச்சுச் செயல்பாடுகள் ஆகியவற்றின் அடிப்படையில் புனைகதைப் பனுவல்களையும் கணினி வழி இலக்கிய ஆய்வுக்கு உட்படுத்த இயலும்.

துணைநூல்கள்

- காமாட்சி, அ., கல்பனா, செ., 2016, ஐங்குறுநூறு உருபனியற் பகுப்பாய்வு, நியூ செஞ்சுரி புக் ஹவுஸ் (பி) லிட், சென்னை-98,
- சுசீலா மு., 2003, பழந்தமிழ்த் தொடரியல், தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர்- 5
- டேவிட் பிரபாகர் ப., 2018, சங்கப் பனுவல் - சொல் வகைகளின் வருகையும் பகிர்வும்: புள்ளியியல் நோக்கு, செம்மொழி தமிழாய்வு நிறுவனக் குறுந்திட்ட ஆய்வறிக்கை, சென்னை-100
- நீதிவாணன் ஜெ., 1983, நடையியல், மணிவாசகர் பதிப்பகம், சிதம்பரம்,
- பேராசிரியர் (உரை) 1975, தொல்காப்பியம், பொருளதிகாரம், கழகம் சென்னை-1
- வீரப்பன் பா., 1989, சங்க இலக்கிய நடை, பூவழகி பதிப்பகம், சென்னை 14
- Gius, E. & Vauth, M., 2022, Towards an Event Based Plot Model. A Computational Narratology Approach, Journal of Computational Literary Studies 1. doi: https://doi.org/10.48694/jcls.110
- Giuseppina Balossi 2014 A Corpus Linguistic Approach to Literary Language and Characterization Virginia Woolf's The Waves ,John Benjamins publishing company.
- Hatzel, Hans Ole, Stiemer, Haimo, Biemann, Chris and Gius, Evelyn. "Machine learning in computational literary studies" it - Information Technology, vol. 65, no. 4-5, 2023, pp. 200-217. https://doi.org/10.1515/itit-2023-0041
- Jacobs AM, Kinder A, 2022, Computational analyses of the topics, sentiments, literariness, creativity and beauty of texts in a large Corpus of English Literature, arXiv preprint arXiv:2201.04356, 2022•arxiv.org
- Mark Aronoff, Janie Rees - Miller, 2017, The Handbook of Linguistics, John Wiley & So Tess M. E. A. Crosbie, A COMPUTER ASSISTED ANALYSIS OF LITERARY TEXT: FROM FEATURE ANALYSIS TO JUDGEMENTS OF LITERARY MERIT (Ph.D. Thesis), University of Bedfordshire, November 2016

தற்காலத் தமிழ் மொழிக்கான சொற்பொருட்களஞ்சியம்

கோ.பழனிராஜன்

ஆய்வுச்சுருக்கம்

'Thesaurus' என்ற வார்த்தை கிரேக்க வார்த்தையான 'thesauros' என்பதிலிருந்து வந்தது, அதாவது களஞ்சியம் அல்லது வார்த்தைகளின் கருவூலம் என்று பொருள். சொற்களின் உறவுத்தொகுப்பு சொற்பொருட் களஞ்சியம் (thesaurus) எனப்படும். Thesaurus என்ற ஆங்கிலச் சொல்லுக்கு இணையாகத் தமிழில் சொற்பொருட் களஞ்சியம் என்று இங்கு குறிப்பிடப்படுகிறது. இதைப் பொருட்புல அகராதி என்றும் குறிப்பிடுவர் (இராஜேந்திரன் 2001). கருத்துகளிலிருந்து அல்லது பொருண்மைகளிலிருந்து சொற்களைப் பெற உதவுவது சொற்பொருட் களஞ்சியம். இதைப் பற்றி விளக்கமாகப் பின்வருமாறு ஜோன்ஸின் (Jones, 1986:201): "சொற்பொருட் களஞ்சியத்தில் ஒருபொருள் பலசொல் இருக்க வேண்டிய கட்டாயம் இல்லை. மற்றும் ஒருபொருள் பலசொல் அகராதி சொற்களஞ்சியமாக அமையத் தேவையில்லை". சொற்பொருட் களஞ்சியம் என்பது சொற்களைக் கருத்துருக்கள் (Concepts), தலைப்புகள் (topics) அல்லது பாடப்பொருள்கள் (Subjects) ஆகியவற்றின் அடிப்படையில் பாகுபாடு செய்வதாகும் (இராஜேந்திரன் 2001). அதாவது ஒவ்வொரு சொல்லுக்கும் ஒருபொருட் பலசொல் (Synonymy) இருந்தாலும் இல்லாவிட்டாலும் அது சினைமுதல் உறவுடன் (Part Whole Relation) இணைக்கப்பட்டிருக்க வேண்டும் என்பதை ஜோன்ஸ் வலியுறுத்துகிறார்.

இந்தியாவில் ஆதிகாலத்தில் உரிச்சொல் பயன்பாட்டுமுறை மொழிக் கற்றலில் மிக முக்கியப் பங்குவகித்தது. ஐரோப்பியர் வருக்கைக்குப் பின் அது மெல்லமெல்லக் குறைந்து அகராதியைப் பயன்படுத்தும் முறை பயன்பாட்டிற்கு வந்தது. சொற்களை அகர வரிசையில் அமைத்து அவற்றின் இலக்கணப் பொருளையும் சொற்பொருளையும், சமூக மொழியியல் பொருளையும் ஒரு குடையின்கீழ் அறிந்துகொள்வது மிகவும் எளிமையாக இருந்ததே இதற்குக் காரணம்.

கோ.பழனிராஜன்
கேரள மத்தியப் பல்கலைக்கழகம், காசர்க்கோடு.
மின்னஞ்சல்: gprajancuk@gmail.com

1. தமிழ் மொழிச் சொற்பொருட் களஞ்சியம்

இன்றைய தமிழுக்குச் சொற்பொருட் களஞ்சியம் அவசியம் தேவை. இதுவரை தற்காலத் தமிழுக்கெனச் சிறப்பாக உருவாக்கவில்லை. தற்காலத் தமிழ்ச் சொற்களைப் பயன்படுத்தி அகராதி (Dictionary) உருவாக்கப்பட்டுள்ளதைப் போன்று சொற்பொருட் களஞ்சியம் உருவாக்க வேண்டும். இராஜேந்திரன் (2001) தற்காலத் தமிழ்ச் சொற்களஞ்சியம் என்று குறிப்பிட்டாலும் அதில் சொற்கள் பழந்தமிழ், இடைத்தமிழ் சொற்களும் காணப்படுகின்றன. எடுத்துக்காட்டாக:

1.1 விண்வெளி பற்றியவை

கா, வானம், ஆகாயம், ஆகாசம், விண், விசும்பு, அந்தரிட்சம், அந்தரம், கீழ்வானம், அடிவானம், மேல்வானம் என்று குறிப்பிட்டுள்ளார். இவற்றுள் விசும்பு, அந்தரிட்சம், அந்தரம் முதலிய மூன்று சொற்களும் தற்காலப் பயன்பாட்டில் இல்லாதவை. கீழ்வானம், அடிவானம், மேல்வானம் முதலிய மூன்று சொற்களும் அதன் வகைகளைக் குறிப்பவை. மேலும் ஆகாசம் என்பது பேச்சு வடிவமாகும்.

இவ்வாறாகத் தற்காலத் தமிழ்ச் சொற்கள் குறைவாகவும், நிகண்டுச் சொற்களும், பழந்தமிழ்ச் சொற்களும் மிகுந்தும் காணப்படுகின்றன. மேலும் அவருடைய சொல்வலை (WordNet) இந்திச் சொல்வலை மூலமொழியாகவும் (source language) தமிழ் மொழி இலக்கு மொழியாகவும் கொண்டு சொற்கள் (Target language) நிரப்பப்பட்டன. ஆகவே தற்காலத் தமிழ் மொழிக்கெனச் சிறப்பான சொற்களஞ்சியம் அமைய வேண்டும். குறிப்பாத் தற்காலத் தமிழ் பொதுச் சிறப்புச் சொற்பொருட் களஞ்சியம் அமைய வேண்டும் என்பது பற்றி இக்கட்டுரை ஆராய்கிறது.

1.2. உரிச்சொல்

தமிழில் முதன்முதலில் தொல்காப்பியத்தில் உரிச்சொல் என்ற சொல் சொற்பொருட் களஞ்சியம் பொருண்மைப் பற்றிப் பேசுகிறது. தொல்காப்பியர் சொல்லதிகாரம் உரியியலில் 120 உரிச்சொற்களைக் குறிப்பிடுகிறார். இந்தச் சொல்லுக்கு இன்ன பொருள் என்றும், இந்தப் பொருளை இன்னின்ன சொற்கள் உணர்த்தும் என்றும் இங்குச் சொல்லப்படுகிறது. இதன் அடிப்படையில் பல நிகண்டு நூல்கள் எழுந்தன. அவற்றைப் பின்வரும் பகுதியில் பார்ப்போம்.

2. நிகண்டுகள்

காலத்தால் முந்தியது சேந்தன் திவாகரம் நிகண்டு. திவாகரன் முதற்கொண்டு ஏறத்தாழ 80 நிகண்டுகள் தமிழில் தோன்றியுள்ளன. (சற்குணம் 2002) பொதுவாக மொழி நோக்கில் மட்டுமின்றி சிறப்பாக மருத்துவம், சோதிடம் போன்ற துறைகளிலும் நிகண்டுகள் இயற்றப்பட்டன. இருப்பினும் மொழியியல் நோக்கில் மட்டும் இதுவரை ஏறத்தாழ 50 நிகண்டுகள் கிடைத்துள்ளன. அவற்றுள் 20 நிகண்டுகள் மட்டுமே பதிப்பிக்கப்பட்டுள்ளன. தற்கால நிகண்டு பற்றி ஆய்வுக்கு முன்னோடியாகவும் வழிகாட்டியாகவும் விளங்கியவர் பேராசிரியர் எஸ்.வையாபுரிப்பிள்ளை. தமிழ் அகராதியின் ஆதார நூல் தொகுதி என்னும் தலைப்பின் கீழ் நான்கு நிகண்டுகளை முதல் முதலில் வெளியிட்டார். பேராசிரியர் வ.ஜெயதேவன் தமிழ் அகராதியில் வளர்ச்சி வரலாறு என்னும் தலைப்பில் முனைவர் பட்ட ஆய்வேட்டில் இரண்டு இயல்கள் தமிழ் நிகண்டுகள் தொடர்பான ஆய்வு முதன் முதலில் மேற்கொள்ளப்பட்டது. பேராசிரியர் சுந்தர சண்முகனார் தமிழ் அகராதிக்கலை என்னும் நூலில் இரண்டு, மூன்று பாகங்கள் நிகண்டுகள் பற்றிய செய்திகள் விளக்கமாகத் தரப்பட்டுள்ளன. இன்னாசி சு.சதுரகராதி ஆராய்ச்சி என்ற நூலை நிகண்டுகளுடன் ஒப்பிட்டு ஆராய்ந்துள்ளார். பேராசிரியர் சற்குணம் (2002) தமிழ் மொழியில் நிகண்டுகள் என்ற விரிவான ஆய்வை முதன் முதலில் மேற்கொண்டுள்ளார். அவற்றுள் சில நிகண்டுகளைச் சுருக்கமாகக் காண்போம்.

i. திவாகர நிகண்டு

நிகண்டுகளுள் மிகவும் பழமையானது திவாகர நிகண்டு எனப்படும். சேந்தன் திவாகரம் நிகண்டின் சொற்றொகைப் பாகுபாடு மிகவும் சிறப்பு வாய்ந்தது. ஏனெனில் பிற்கால நிகண்டுகளுக்கு அடிப்படையானது. இது சொற்களைப் பன்னிரண்டு தொகுதிகளாகப் பிரித்து அடக்கியுள்ளது. அவை கீழே கொடுக்கப்பட்டுள்ளன.

1. தெய்வப் பெயர்த் தொகுதி, 2. மக்கட் பெயர்த் தொகுதி, 3. விலங்கினப் பெயர்த் தொகுதி, 4. மரப் பெயர்த் தொகுதி, 5. இடப் பெயர்த் தொகுதி, 6. பல்பொருட் பெயர்த் தொகுதி, 7. செயற்கை வடிவப் பெயர்த் தொகுதி, 8. பண்பு பற்றிய பெயர்த் தொகுதி, 9. செயல் பற்றிய பெயர்த் தொகுதி, 10. ஒலி பற்றிய பெயர்த் தொகுதி, 11. ஒருசொல் பல்பொருள் பெயர்த் தொகுதி, 12. பல்பொருள் கூட்டத்து ஒருபெயர்த் தொகுதி எனப்படும்.

ii. பிங்கல நிகண்டு

பிங்கலர் இயற்றியது. இவர் திவாகரன் மகன் என்று சிறப்புப் பாயிரத்தால் அறிய முடிகிறது. இந்நூல் பத்துத் தொகுதிகளுடன் பிதினைந்து ஆயிரத்து எண்ணூறு சொற்களை உடையது. 1. வான் வகை, 2. வானவர் வகை, 3. ஐயர் வகை, 4. அவனி வகை, 5. ஆடவர் வகை, 6. அனுபொக வகை, 7, பண்பின் செயலின் வகை, 8. மாப்பெயர் வகை, 9. மரப்பெயர் வகை, 10. ஒருசொல் பல்பொருள் வகை. இப் பத்து வகைகளுள் முதல் ஒன்பதும் ஒருபொருள் பல்பெயர்த் தொகுதியாகும்.

iii. உரிச்சொல் நிகண்டு

காங்கேயர் இயற்றியது. மூவாயிரத்து இருநூறு சொற்களை உடையது.

iv. கயாதர நிகண்டு

கயாதரன் இயற்றியது. பதினொரு தொகுதிகளும். பத்தாயிரத்து ஐநூறு சொற்களை உடையது.

v. பாரதி தீபம்

திருவேங்கட பாரதி இயற்றியது. பன்னிரண்டு தொகுதிகளும் பதிநான்காயிரத்து எழுநூறு சொற்களையும் கொண்டது.

vi. சூடாமணி நிகண்டு

மண்டல புருடர் இயற்றியது. பன்னிரண்டு தொகுதிகளும். ஆயிரத்து ஐநூற்று எழுபத்து ஐந்து சொற்களை உடையது.

vii. அகராதி நிகண்டு

புலியூர்ச் சிதம்பர ரேவண சித்தர் இயற்றியது. பொருண்மையின் எண்ணிக்கை அடிப்படையில் சொற்களையும் கொண்டது.

viii. ஆசிரிய நிகண்டு

பத்து தொகுதிகளும். பன்னிரண்டாயிரம் சொற்களை உடையது.

ix. பல்பொருள் சூடாமணி

ஈசுரபாரதி இயற்றியது. அமரகோசத்தைப் பின்பற்றி எழுதப்பட்டது

x. கைலாச நிகண்டு

கைலாசம் இயற்றியது. ஐம்பத்து பிரிவுகளும் பதினைந்தாயிரம் சொற்களும் உடையது.

xi. அகராதி மோனை அகராதி எதுகை

ஏழாயிரத்து ஐந்நூறு சொற்கள் உள்ளன.

xii. அரும்பொருள் விளக்க நிகண்டு

மூவாயிரத்து இருநூறு சொற்கள் உள்ளன.

xiii. பொருட்டொகை நிகண்டு

பத்தாயிரம் சொற்கள் உள்ளன.

xiv. பொதிகை நிகண்டு

சாமிநாத கவிராயர் இயற்றியது.

xv. உசித சூடாமணி நிகண்டு

பதினெட்டுப் பிரிவுகள் கொண்டது.

xvi. நாமதீப நிகண்டு

நான்கு படங்கள் பன்னிராண்டாயிரம் சொற்கள் உள்ளன.

xvii. வேதகிரியார் சூடாமணி நிகண்டு

இரண்டாயிரத்து ஐந்நூற்று இருபத்து ஆறு சொற்கள் உள்ளன.

xviii. தொகைப் பெயர் விளக்கம்

களத்தூர் வேதகிரி முதலியார் இயற்றியது.

xix. கந்தசுவாமியம்

சுப்பிரமணிய தேசிகர் இயற்றியது.

xx. இலக்கத் திறவுகோல்

தொகைப் பெயர்கள் குறிக்கப்பட்டுள்ளன.

இவ்வாறாகச் சில நிகண்டுகள் கிடைத்துள்ளன. பல நிகண்டுகளின் பெயர்கள் மட்டுமே கிடைத்துள்ளன. அவற்றுள் 20 நிகண்டுகள் மட்டுமே பதிப்பிக்கப்பட்டுள்ளன என்று முன்பே குறிப்பிட்டுள்ளார். இவற்றை அரசு நிறுவனங்கள் பொறுப்பேற்றுப் பதிப்பித்துப் பாதுகாக்க வேண்டும்.

3. Roget's Thesaurus (1852)

ரோகேட் சொற்பொருட் களஞ்சியம் பத்து (10) முதன்மையான சொற்றொகைப் பாகுபாட்டையும் எண்ணூற்று முப்பத்து ஏழு (837) துணைப் பாகுபாட்டையும் கொண்டது.

1. செயல் (Actions), 2. காரணம் (Causes), 3. மனிதச் செயல்கள் (Fields of Human Activity) 4. வாழ்க்கை வளர்ச்சி நிலை (Life Forms) 5. பொருள்கள் (Objects,) 6. விண்வெளி (The Planet), 7. பண்புகள் (Qualities), 8. உணர்வுகள் (Sense), 9. நிலைகள் (States), 10, எடை, அளவுகள் (Weights and Measures)

4. இராஜேந்திரன் (2001)

இராஜேந்திரன் சொற்பொருட் களஞ்சியம் நான்கு (4) முதன்மையான சொற்றொகைப் பாகுபாட்டை கொண்டது.

பருப்பொருள்கள் (objects), 2. நிகழ்வுகள் (events), 3. அருவங்கள் (abstracts), 4. தொடர்பன்கள் (relations) முதன்மை சொற்றொகை வகைப்பாட்டையும்

தற்காலச் சொற்பொருட் களஞ்சியம் (Thesaurus)

சொற்பொருட் களஞ்சியம் கருத்துருக்களிலிருந்து அல்லது பொருண்மைகளிலிருந்து சொற்களைப் பெற உதவும். இதை இரு பெரும் பிரிவுகளாகப் பார்க்கலாம்.

அ) ஒரு குறிப்பிட்ட புலத்தின் சொற்கள் அல்லது ஒத்த கருத்துகளின் தொகுப்பு பற்றிய இணைச்சொற்கள் சொற்பொருட் களஞ்சியம் எனலாம்.

ஆ) ஒருபொருள் பலசொற்கள், பொருட்புலச் சொற்களுடன் படிநிலைச் சொற்களையும் கருத்துறவுடன் (cross-reference system) அறிந்துகொள்ள உதவுவது .

4.1 சொற்பொருட் களஞ்சியம் கட்டமைப்பு (Thesaurus Construction)

சொற்பொருட் களஞ்சியத்தை வடிவமைப்பது என்பது திட்டமிடல், விரிவாக்கம், செம்மையாக்கம் ஆகிய பல கட்டங்களை உள்ளடக்கிய ஒரு சிக்கலான பணியாகும். ஒருபொருள் பலசொற்கள் (synonyms), எதிர்ச்சொற்கள் (antonyms) மேலும் பல்வேறு வழியில் தொடர்புடைய சொற்களைக் (related in some other words) கண்டறிதல் இதன் நோக்கமாகும். சொற்பொருட் களஞ்சியத்தின் கட்டமைப்பைக் கீழ்க்கண்ட படிநிலைகளில் காணலாம்.

4.2 சொற்பொருட் களஞ்சிய வகைகள் (Types of Thesaurus)

பல்வேறு வகைகளில் சொற்பொருட் களஞ்சியத்தை வடிவமைக்கலாம். சொற்பொருட் களஞ்சியம் ஒவ்வொன்றும் வெவ்வேறு நோக்கத்தையும் இலக்கையும் அடிப்படையாகக் கொண்டது. ஒவ்வொரு வகையிலும் கட்டமைப்பு, உள்ளடக்கம் மற்றும் தரவுகளின் நிலை கணிசமாக வேறுபடலாம். இங்கே சில பொதுவான வகைகளைப் பற்றிக் காண்போம்.

4.3 பொது மொழி சொற்பொருட்க ளஞ்சியம் (General Language Thesaurus)

ஒருபொருள் பலசொற்கள் சொற்பொருட் களஞ்சியம்: இது பொதுவாகப் பயன்படுத்தப்படும் வகை. இது சொற்களை அவற்றின் ஒத்த சொற்கள் மற்றும் பெரும்பாலும் அவற்றின் எதிர்ச்சொற்களுடன் பட்டியலிடுகிறது. எடு. ரோஜெட்டின் தெசரஸ் (Roget's Thesaurus) மற்றும் மெரியம்-வெப்ஸ்டர் தெசரஸ் (Merriam-Webster Thesaurus) ஆகியவை அடங்கும்.

4.4 கருத்தியல் சொற்பொருட் களஞ்சியம் (Conceptual Thesaurus)

சொற்பொருட் களஞ்சியத்தில் ஒருபொருள் பலசொற்கள் (synonymy) மட்டுமல்லாமல், தொடர்புடைய கருத்துகளாலும் (concepts) ஒருங்கிணைத்து உருவாக்குவது. இது “சமையல் தொடர்பான சொற்கள்” அல்லது “நேரம் தொடர்பான சொற்கள்” போன்ற பல்வகையான சொற்களை உள்ளடக்கியது. எடுத்துக்காட்டாக;

எதிர்ச்சொல்

எதிர்ச்சொல் (Antonym)

துருவ எதிர்ச்சொல் (Polar antonym) – குள்ளம் உயரம்

எதிர் எதிர்ச்சொல் (overlapping antonym) – நல்லது–கெட்டது

தொடர்பு எதிர்ச்சொல் (complementarily antonym) திற மூடு

மறுதலை (conversances) கணவன் மனைவி

உறவுமுறை (relation) மகன் மகள்

காலம் (Temporal relations) முன்னே பின்னே

இடம் (Spatial relation) முன்னே பின்னே

4.5 கட்புலப் சொற்பொருட் களஞ்சியம் (Visual Thesaurus)

வார்த்தைகளுக்கு இடையே உள்ள உறவுகளை வரைபடமாக வெளிக்கொணர்வது. சொற்களுக்கு இடையே உள்ள கோடுகள் அல்லது அம்புகள் ஒத்த சொற்கள், எதிர்ச்சொற்கள் மேலும் பரந்த/குறுகிய சொற்கள் போன்ற உறவுகளின் வகைகளைக் குறிக்கின்றன.

4.6 வரலாற்றுப் சொற்பொருட் களஞ்சியம் (Historical Thesaurus)

இந்த வகை பல்வேறு காலங்களின் சொற்களை உள்ளடக்கியது, அவை எப்போது பயன்பாட்டில் இருந்தன என்பதைக் குறிக்கிறது. வரலாற்று மொழியியலில் ஆர்வமுள்ள அறிஞர்களுக்கு மிகவும் பயனுள்ளதாகும்.

4.7 மரபுத்தொடர் சொற்பொருட் களஞ்சியம் (Idiomatic Thesaurus)

சொற்கள், தொடர்கள் இணைந்து புதிய பொருண்மையை உணர்த்துவது.

4.8 சிறப்புப் சொற்பொருட் களஞ்சியம் (Specialized Thesaurus)

சிறப்புப் புலப் சொற்பொருட் களஞ்சியம் (Domain-Specific Thesaurus): மருத்துவம், பொறியியல் போன்ற ஒரு குறிப்பிட்ட துறைக்காகச் சிறப்புப் புலப் சொற்பொருட் களஞ்சியம் உருவாக்கப்படுவன. இவை மிகவும் துல்லியமானவை மற்றும் பொதுவாகப் பொது சொற்களஞ்சியத்தில் காணப்படாத சொற்கள், தொடர்கள் உள்ளடக்கியிருக்கும்.

4.9 பன்மொழி சொற்பொருட் களஞ்சியம் (Multilingual Thesaurus)

இது பல மொழிகளின் சொற்களை உள்ளடக்கியது, பெரும்பாலும் ஒரு மொழியின் ஒருபொருள் பலசொற்கள் இணையாக இலக்கு மொழியின் இணை/ ஒத்த/ சமமான சொற்களை இட்டு நிரப்புவதாகும்.

4.10 வட்டாரப் சொற்பொருட் களஞ்சியம் (Regional Thesaurus)

ஒரு குறிப்பிட்ட நிலப்பகுதியில் ஒரு குறிப்பிட்ட பேச்சுவழக்குகளைத் தொகுத்து உருவாக்குவது.

4.11 குழந்தைகளுக்கான சொற்பொருட் களஞ்சியம் (Children's Thesaurus)

இது பெரும்பாலும் எளிமையான சொற்களைக் கொண்டு உருவாக்குவது. சொற்களுக்கு எளிமையான விளக்கங்களும் விளக்கப்படங்களும் கொடுக்கப்பட்டிருக்கும்.

4.12 கல்விசார் சொற்பொருட் களஞ்சியம் (Academic Thesaurus)

கல்வியியல் சார்ந்த சொற்களை இலக்காகக் கொண்டு உருவாக்குவது, இந்த வகை மிகவும் சிக்கலான சொற்கள் மற்றும் தொழில்நுட்பச் சொற்களை உள்ளடக்கியிருக்கும்.

5. இலக்கமுறை சொற்பொருட் களஞ்சியம் (Digital Thesaurus)

இலக்கமுறை சொற்பொருட் களஞ்சியம் என்பது மேற்கண்ட பன்னிரண்டு சொற்பொருட் களஞ்சியதையும் இலக்கமுறை சொற்பொருட் களஞ்சியமாக மாற்ற முடியும். அவற்றுள்ளும் மூன்று பெரும் பிரிவுகளாகப் பிரிக்கலாம். அவை; 1. சொற்பொருட் களஞ்சியம் (Online Thesaurus), இயங்கு சொற்பொருட் களஞ்சியம் (Dynamic Thesaurus), 3. ஊடாட்டுச் சொற்பொருட் களஞ்சியம் (Interactive Thesaurus)

5.1 சொற்பொருட் களஞ்சியம் (Online Thesaurus)

இது இணைய சொற்பொருட் களஞ்சியம் எனப்படும். பெரும்பாலும் இணைய செயலிகளில் (Apps) ஒருங்கிணைக்கப்பட்டது. இது தரவுகள், ஒலிக்குறிப்புகள், விளக்கப்படங்கள் போன்ற பல்வேறு சிறப்புத் தகவல்களுடன் மேம்படுத்தப்பட்ட சொற்பொருட் களஞ்சியமாக விளங்குகிறது.

5.2 இயங்கு சொற்பொருட் களஞ்சியம் (Dynamic Thesaurus)

புதிய சொல் பயன்பாட்டுச் சூழலில் பல்வேறு புதிய சொறகளைத் தரவகத்திலிருந்து தானியங்கி நிகழ்நிலை முறையில் புதுப்பித்துக்கொள்கின்றது. பெரும்பாலும் இயந்திர கற்றல் அல்லது சொற்குவியல் (crowd sourcing) ஆதாரமாகப் பயன்படுத்துகின்றன.

5.3 ஊடாட்டுச் சொற்பொருட் களஞ்சியம் (Interactive Thesaurus)

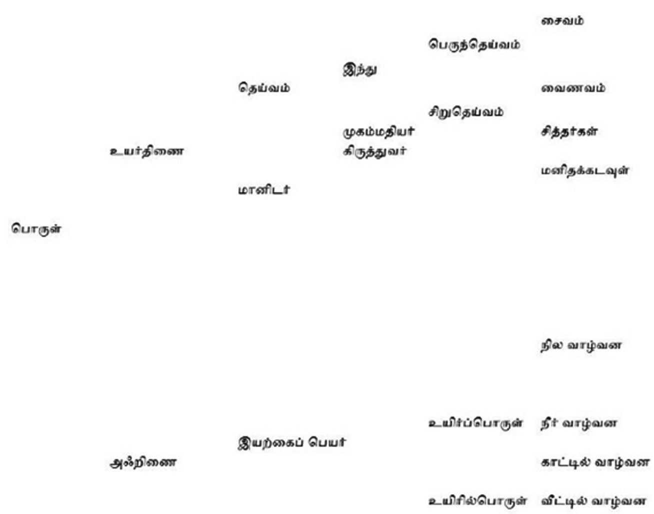
பயனர்கள் தங்கள் சொந்த ஒருபொருள் பலசொற்களைச் சேர்க்க அனுமதிக்கிறது, இது பயனர்

ஈடுபாட்டை வளர்க்கவும், அறிவாற்றலை மேம்படுத்தவும் வழிவகுக்கிறது. மேலும் சொற்களின் உறவுகளைத் தேர்வு செய்து ஓட்டுப்போடவும் இயலும்.

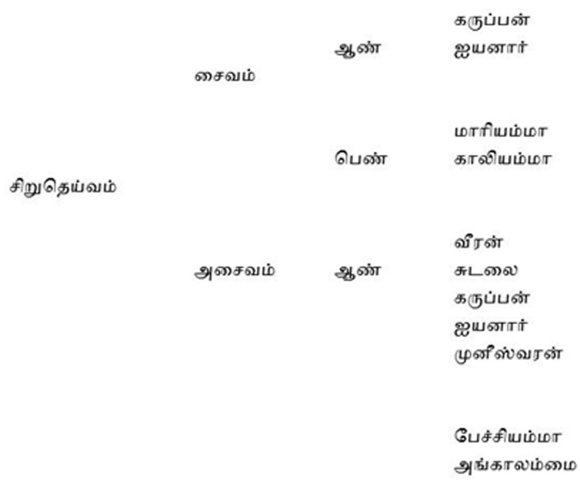
6. சொற்பொருள் உறவுகள் (Semantic Relationships):

சொற்பொருள் உறவுகளின் அடிப்படையில் மீச்சொல் (hypernymy) உயர் உள்ளீடு (hyponym) சினைமுதல் உறவு (meronymy) போன்ற பிற வகையான உறவுகளை அடையாளம் காண்டு சேர்க்க வேண்டும். எடுத்துக்காட்டாக நாற்காலி என்ற சொல்லின் சொற்பொருள் உறவுகளைக் காண்போம்.

பொருள்→ அஃறிணை→ செயற்கை→ நிலம்→ வீட்டு வீட்டு உபயோகப்பொருள்→ மரச்சாமன்கள்→ நாற்காலி.



படம் 1. சொற்பொருட் களஞ்சியம் மாதிரி



படம் 2. சொற்பொருட் களஞ்சியம் மாதிரி

<பொருள்

<அஃறிணை

<செயற்கை

<நிலம்

>வீட்டு உபயோகப்பொருள்

<மரச்சாமன்கள்

<நாற்காலி

முடிவுரை

இன்றைய தமிழுக்குச் சொற்பொருட் களஞ்சியம் அவசியம் தேவை. தற்காலத் தமிழ்ச் சொற்களைப் படுத்திச் சொற்பொருட் களஞ்சியம் உருவாக்க வேண்டும். நிகண்டுகளுள் மிகவும் பழமையானது திவாகர நிகண்டு எனப்படும். திவாகர நிகண்டின் சொற்றொகைப் பாகுபாடு மிகவும் சிறப்பு வாய்ந்தது ஏனெனில் பிற்கால நிகண்டுகளுக்கு அடிப்படையானது. இது சொற்களைப் பன்னிரண்டு தொகுதிகளாகப் பிரித்து அடக்கியுள்ளது. இந்த அமைப்புமுறையுடன் ஆங்கிலச் சொற்பொருட் களஞ்சியத்தையும் ஒருங்கிணைத்து தமிழுக்கு உருவாக்க வேண்டும். அது எவ்வாறு அமைக்க வேண்டும் என்பது பற்றி திட்ட வரைவு கொடுக்கப்பட்டுள்ளது. அதை அடிப்படையாகக் கொண்டு உருவாக்க வேண்டும். தற்கால அகராதியில் இருந்து சொற்பொருட் களஞ்சியத்தை வடிவமைப்புது என்பது மிகவும் சிக்கலான பணியாகும். இருந்தபோதிலும் சொற்களை ஒருபொருள் பலசொற்கள் (synonyms), எதிர்ச்சொற்கள் (antonyms) மேலும் பல்வேறு வழியில் தொடர்புடைய சொற்களைக் (related in some other words) மீச்சொல் வகையோடு இணைக்கவேண்டும். பல்வேறு வகைகளில் சொற்பொருட் களஞ்சியத்தை வடிவமைக்கலாம் என்பது பற்றியும் விரிவாக இங்கு பேசப்பட்டுள்ளது. சொற்பொருட் களஞ்சியம் ஒவ்வொன்றும் வெவ்வேறு நோக்கத்தையும் இலக்கையும் அடிப்படையாகக் கொண்டது. ஒவ்வொரு வகையிலும் கட்டமைப்பு, உள்ளடக்கம் மற்றும் தரவுகளின் நிலைகளின் அடிப்படையில் உருவாக்கவேண்டும் என்பதை இவ் ஆய்வு வலியுறுத்துகிறது.

துணைநூல்கள்

- இராசேந்திரன், ச. 2001. தற்காலத் தமிழ்ச் சொற்களஞ்சியம். தமிழ்ப் பல்கலைக்கழகம், தஞ்சாவூர்.
- சற்குணம்,மா (2002) தமிழ் நிகண்டுகள் ஆய்வு சென்னை; இலவழகன் பதிப்பகம்.
- Cruse,D.A. 1986. Lexical Semantics. Cambridge: Cambridge University Press.
- Cruse, D.A. 2000. Meaning in Language: An Introduction to Semantics and Pragmatics. Oxford: Oxford University Press.
- Karunakaran, K. 1987. Langue Planning in Tamil: retrospects and prospecs. Tamil Civilization, vol 5, no. 3, 58-65.
- Leech, Geoffery .1981. Semntics, Penguin
- Leech, Geoffery .1983. Principles of Pragmatics, London: Longman.
- Lyons, J. 1963. Structural semantics. Oxford: Blackwell.
- Lyones, J. 1977. Semantics (vol.) Cambridge: Cambridge University Press.
- Mawson, C.O.S. 1956. Roget's International Thesaurus of English Words and Phrases. New York: Pocket Books, INC.
- Nida, E.A. 1975a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton.
- Nida, E.A.. 1975.b. Exploring Semantic Structure. The Hague: Mouton
- Rajendran, S. 1978. Syntax and Semantics of Tamil Verbs. Ph.D. Thesis. Poona: University of Poona.
- Sundarabalu, S. Ed,. 2021. A meaning centric tool for making thesaurus in Tribal Languages. Coimbatore: Bharathiar University.

Sign language model for Hearing Impaired People using LLM

R. Krithiga, S. Shoba

ABSTRACT

Hearing-impaired people have the primary challenge of communicating with normal people. The main goal is to make it possible for people with disabilities and robots to communicate without speech. Without the conversation, this work quickly recognizes the hand gestures and displays the text in Tamil language. This work focuses on recognizing the real-time activity words using Media Pipe process with EfficientNetB1 to understand hand gestures in real time. We created a own dataset named TActSign of 3000 samples which contains 10 activity words. The recognized gestures produce significant results and compared with the state of art methods.

R. Krithiga

School computer science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India.

S. Shoba

Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, Tamil Nadu, India.

INTRODUCTION

Indian sign language (ISL) is a native and natural language used by the deaf community. It works on the visual language to communicate with the deaf people. ISL has gained recognition as a separate language, and efforts are being made to standardize it and promote its use. In 2019, the Indian government recognized ISL as a linguistic minority language and included it in the Eighth Schedule of the Indian Constitution. ISL has its own unique vocabulary and grammar. It relies on hand movements, facial expressions, body postures, and other visual cues to convey meaning. These elements are structured in a specific way to form sentences and convey complex ideas. ISL is the only way for the hearing impaired people for communication. The Government has failed to provide an appropriate free public education to the disabilities and the special educators in this community are less in number. Only few number of impaired people are educated and attending the impaired programmes. Special education is lacking among the states that falls under the category of low resource languages (LRL). Across India, the highest number of disabled has been reported from the state of Uttar Pradesh (3.6 million). Significant numbers of disabled have also been reported from the state like Bihar (1.9 million), West Bengal (1.8million), Tamil Nadu and Maharashtra (1.6 million each). Among the states, Arunachal Pradesh has the disabled males (66.6%) and lowest proportion of female disabled. Tamil Nadu is the only state, which has a higher number of disabled females than males.

With the advancement of technologies such as Machine Learning, Deep Learning, Computer Vision Sign Language Recognition (SLR) systems have been made possible. These recognition systems are the solution to these challenges. Sign Language Recognition is an advanced technology to aid the communication process. The goal of Sign Language recognition systems is to translate signs into understandable formats such as text or speech. Sign language involves the use of Hand gestures as means of communication. The Challenges faced by deaf students are as follows:

- The normal people cannot understand the signs delivered by the deaf peoples

- Difficulty in communication between the normal and hearing impaired people
- Without educators, the exact ISL cannot be taught to the deaf peoples
- Inaccessibility of education for deaf students is crucial.

RELATED WORK

The LRL in the field of natural language processing (NLP) has been established by researchers based on the presence of data, including labeled, unlabelled, or auxiliary data, as well as the availability of NLP tools and resources [1]. The LRL is a growing field with numerous challenges and difficulties. Techniques widely used in ISL recognition, are Convolution Neural Networks, Gaussian Filtering, Speech-to-text, Video-to-text. As the availability of dataset is less (LRL), researchers establish many findings by examining the signs of native and mother tongue speakers. A library named open hands [2] that leverages four fundamental concepts in LRL of word level recognition. This work creates a poses of various word sign recognition symbols for six languages namely, American, Chinese, Indian, Greek, Turkish and Argentina. This work trains the model efficiently with less time using pose extracted pre-trained models.

A real-time two-way sign language communication system [9] built using image processing, deep learning and computer vision. CNN model trained with a large dataset for 40 classes and was able to predict 17600 test images in 14 seconds with an accuracy of 99%. [10]. A HMM-based framework [11] used motion sensor gloves or colored wristbands to recognize SLR. Most of the work used glove based method may be accurate but costlier which may not be affordable for rural peoples. The cost effective method called fingerspelled sign learning is the primary phase of sign language acquisition.

Oliveira et al. [12] conducted a comparison between Principal Component Analysis (PCA) and Convolutional Neural Network (CNN) based methods for recognizing fingerspelled letters in Irish Sign Language. The PCA model achieved a recognition accuracy of 0.95, whereas the CNN model achieved a recognition accuracy of 0.99. SignQuiz, is a cost-efficient online tool [13] for learning fingerspelled signs in ISL that utilizes an automated technique for recognizing sign language. The complex methodologies are inadequate in handling large data and are distinguished by processing the image and acquiring valuable information by learning efficacy. It is important to note that the categorization of the Indian Sign Language (ISL) using the Mediapipe API is quicker than traditional approaches and surpasses them in computational ability. A novel MediaPipe-optimized gated recurrent unit (MOPGRU) model [14] designed to recognizing ISL. This work enhances the gate of general GRU cell by performance computational multiplication by eliminating the unnecessary information.

The ISL uses Mediapipe Hands API [Chakraborty, Subhalaxmi] by Google categorizes English alphabets. It uses 21 points in each hand with x, y, z coordinates in 3D space. The recognized ISL was compared with machine learning techniques and achieved significant results.

India is a diversified with large number of languages and people. Each region of the state is observed with different sign languages. Due to the variability in the signs in different state, communication among the hearing impaired between each state is a challenging one. This work focuses on the specific state called Tamil Nadu where preferred language for the people is Tamil. Any sign language could be effectively recognized when an intelligent algorithm is paired with the results of image processing. The approach [3] entails identifying hand movement, monitoring the hand's position based on the movement, and categorizing indications by adaptive clustering of stopping points, the basic shape of the hand's path, and matching the hand's shape at the stopping point.

So a special attention has to be given for producing a sign language dataset for each state in India. Few of the works carried out with different languages like Kannada, Malayalam, and Telugu. In order to generate matching letters in Kannada language, [4] uses curvilinear tracing approach for shape representation and recognition of Kannada sign language. The Kannada dataset is created by defining a vocabulary of different sign symbols. Another work [5] conveys the information in variety of representations, such as articulations, finger signs, and a both. The framework has been developed for analyzing the activity of sign, recognizing and finding the age. The words associated stored in the dataset is tested with five videos and represented in Kannada language in the form of text.

The work [6] develops a prototype which feds text as input and produces output in Malayalam sign language using automatic sign language translator system. The representation of signs can be generated using 3D character animation using the proposed system HamNoSys [7]. This input enters as single word or several words or can add new terms to the database by sign editor that adds into the HamNoSys framework. The output is generated as an animation by the prototype model. An unified finger spelling alphabet Malayalam language dataset for applying to the deep learning model. The model uses transfer learning approach which recognises the alphabetic letters using ResNet50 [8].

Converting Tamil Sign Language Alphabets (TSL) into speech displays a set of 32 pictures representing alphabet of Tamil signs with an accuracy of 99.35% for static and 98.36% for dynamic using angular-based analysis [16]. Another work uses external webcam to capture the test image and detects hand gestures using singular value decomposition and background subtraction [17]. The above research papers focuses on ISL and few on TSL alphabets. As per our knowledge, no work focuses on the words and in particular on the activity signs. This research work not only focuses on collecting the dataset for Tamil language and in addition recognizes the activity sign accurately using an efficient deep neural network

The overview of the work is to capture the sign from the hearing impaired people which removes the background through MediaPipe Framework. The patterns are recognized by applying the features through the EfficientNetB4 model by displaying the output as text.

MATERIALS AND METHODS

Objective

- To create a TActSign dataset for the activity words related to day to day activities.
- To develop the sign language technology by innovative tools and applications for the benefit of each impaired people.
- To assist the individuals with hearing impaired to understand sign language in displayed text.
- To evaluate the activity words through the performance accuracy.

The SLR System created using TAct Sign dataset is to the understand the gestures displayed from hearing impaired people to the common individual. The results are recognized accurately and facilitate the effective communication for individuals with hearing impairments and speech disabilities. Figure 1 shows the overall system architecture of the developed model.

Pre-processing is a primary step for cropping the captured data by removing the background noise and focuses only on the gesture for applying to neural network. The created dataset TActSign was developed after the processing of MediaPipe framework which does the pre-processing steps.

The architecture used in this work is the EfficientNetB4 which applies TensorFlow framework to fine-tune an model that has already been trained on ImageNet to classify faces. The model involves compound scaling, where the depth, width, and resolution of the network are scaled simultaneously. This convolutional approach encompasses two distinct operations: depthwise convolution and pointwise convolution. The depthwise convolution independently filters each input channel spatially, while the subsequent pointwise convolution amalgamates information across channels. Mathematically, the FLOPs for a single layer of depth wise separable convolution can be articulated as:

FLOPsdepthwise separable = (Ks ×Ks × Ic × Sd) + (Ic x Oc × Sd)

For depthwise convolution component

The initial term, (Ks × Ks × Ic × Sd), signifies the FLOPs associated with depthwise convolution

For pointwise convolution component:

The subsequent term, (Ic × Oc × Sd), encapsulates the FLOPs contributed by pointwise convolution. where,

- Ks × Ks - convolutional kernel size
- Ic - number of input channels
- Oc -number of output channels/filters.
- Sd Input feature map spatial dimensions

The adoption of depth wise separability results in a substantial reduction in both parameters and computational load compared to traditional convolutional approaches, making it particularly well-suited for real-time applications like sign language recognition. The efficacy of depth wise separable convolutions lies in their inherent introduction of parallelism and reduction of redundancy in computations, leading to expedited inference times. When considering the overall model, the cumulative FLOPs are obtained by summing the FLOPs across all layers employing depthwise separable convolutions. This reduction in FLOPs positions EfficientNetB4 with depthwise separability as an optimal choice for sign language recognition systems, ensuring that computational requirements are minimized without compromising the model's precision in interpreting and identifying hand gestures as shown in Table 1. The time complexity of computation can be reduced by using depthwise separability and makes the system efficient way of recognition. The power of depthwise separable convolutions comes from their natural use of more than one operation at the same time and less repetition in calculations, making them faster to work. When we look at the whole plan, we add up FLOPs for each layer that uses depthwise separable convolutions. The main intent of these techniques was to increase accessibility and communication for people who have hearing loss. They provided a range of methods for communicating and understanding sign language in different settings.

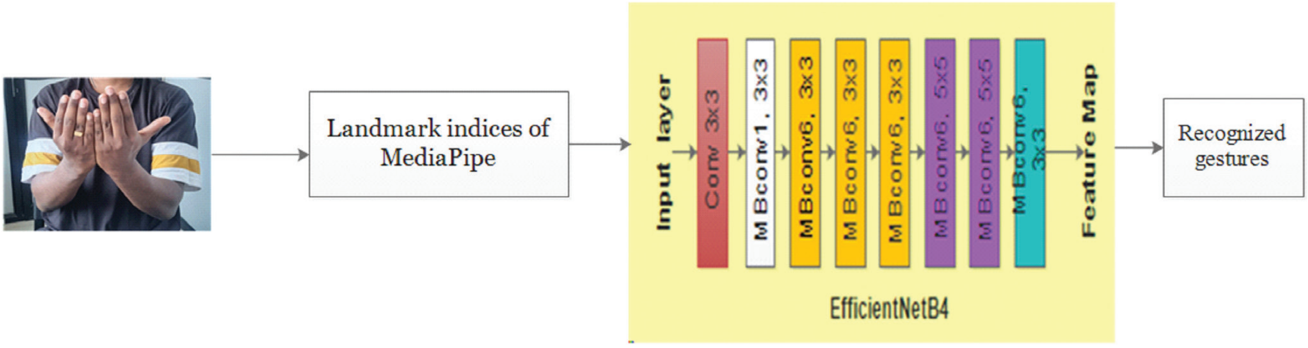


Figure1 System architecture

EXPERIMENTAL RESULTS

The SLR system assesses the findings and examined with the Tact Sign Dataset created. The dataset uses 10 different daily activity signs of both female and male candidates. Each individual signs were captured for 300 images under various lightning conditions and different angles. The total size of the TAct Sign is 3000 samples. The system setup consists of an Intel CPU from the 10th generation i5 series, together with 8GB of RAM. The system is evaluated using metrics such as classification accuracy, precision and recall [18].Table 1 illustrates the evaluation of the performance metric.

Table I. Evaluation metric of a proposed model

S.no	Images	Test accuracy
1	 பசி- Hungry	95.6
2	 பள்ளி- School	92.7
3	 நண்பர்- Friend	93.5
4	 வீடு- House	91.8
5	 கடிகாரம்-clock	93.2
6	 ஆசீர்வாதம்-Blessing	94.6

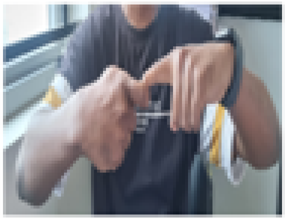
7	 கேள்வி-question	91.3
8	 Answer- பதில்	91.6
9	 தூங்கு-sleep	93.2
10	 காளான்-mushroom	93.8

TABLE 2. Test accuracy for random words.

Table 2 predicts 10 different signs and its test accuracy. Through the observation of the signs denoted in Table 2, the sign hand gestures such hungry, blessing shows a accurate recognition as the co-ordinate points by Mediapipe framework is less and trained in an efficient way. Whereas the recognition such as Friend, Clock, sleeps, mushroom is comparatively better than the single hand as the points co-ordinated and connected the points sequentially. The other gestures shows a less accuracy because of the missing connected points due to the action word of the gesture looks complex than the previous one. The overall average accuracy of the model is 93.15% which shows significant results.

CONCLUSION

The SLR system translates the sign language into the corresponding text with EfficientNetB0+LSTM model. The recognized gesture displayed in Tamil text developed a effective communication between the common individual and the hearing impaired people, A reasonable created TAct Sign dataset of 3000 in number achieved an average accuracy of 93.85% using developed SLR system. In a long run, the dataset can be extended by creating a large of activity words by applying to the generative AI model for the recognition.

REFERENCES

[1] Hedderich, M.A., Lange, L., Adel, H., Strötgen, J. and Klakow, D., 2020. A survey on recent approaches for natural language processing in low-resource scenarios. arXiv preprint arXiv:2010.12309.

[2] Selvaraj, P., Nc, G., Kumar, P. and Khapra, M., 2021. OpenHands: Making sign language recognition accessible with pose-based pretrained models across languages. arXiv preprint arXiv:2110.05877.

[3] Charayaphan, C. and Marble, A.E., 1992. Image processing system for interpreting motion in American Sign Language. Journal of biomedical engineering, 14(5), pp.419-425

[4] Kagalkar, R.M. and Gumaste, S.V., 2019. Curvilinear tracing approach for recognition of Kannada sign language. International Journal of Computer Applications in Technology, 59(1), pp.21-30.

[5] Kumar, A. and Narasimhayya, B.E., 2022, December. An Analysis on Modelling of Kannada Sign language Translator. In 2022 Fourth International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT) (pp. 1-6). IEEE.

[6] Joy, J. and Balakrishnan, K., 2014. A prototype Malayalam to sign language automatic translator. arXiv preprint arXiv:1412.7415.

[7] Nair, M.S., Nimitha, A.P. and Idicula, S.M., 2016, September. Conversion of Malayalam text to Indian sign language using synthetic animation. In 2016 International Conference on Next Generation Intelligent Systems (ICNGIS) (pp. 1-4). IEEE.

[8] Salim, A., 2023, April. Sign language recognition of Malayalam alphabets using transfer learning. In 2023 International Conference on Power, Instrumentation, Control

and Computing (PICC) (pp. 1-4). IEEE.

[9] Bohra, T., Sompura, S., Parekh, K. and Raut, P., 2019, November. Real-time two way communication system for speech and hearing impaired using computer vision and deep learning. In 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 734-739). IEEE.

[10] Cui, R., Liu, H. and Zhang, C., 2019. A deep neural framework for continuous sign language recognition by iterative training. IEEE Transactions on Multimedia, 21(7), pp.1880-1891.

[11] Elpeltagy, M., Abdelwahab, M., Hussein, M.E., Shoukry, A., Shoala, A. and Galal, M., 2018. Multi-modality-based Arabic sign language recognition. IET Computer Vision, 12(7), pp.1031-1039.

[12] Oliveira, M., Chatbri, H., Little, S., Ferstl, Y., O'Connor, N.E. and Sutherland, A., 2017, November. Irish sign language recognition using principal component analysis and convolutional neural networks. In 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA) (pp. 1-8). IEEE.

[13] Joy, J., Balakrishnan, K. and Sreeraj, M., 2019. SignQuiz: a quiz based tool for learning fingerspelled signs in indian sign language using ASLR. IEEE Access, 7, pp.28363-28371.

[14] Subramanian, B., Olimov, B., Naik, S.M., Kim, S., Park, K.H. and Kim, J., 2022. An integrated mediapipe-optimized GRU model for Indian sign language recognition. Scientific Reports, 12(1), p.11964.

[15] Chakraborty, S., Bandyopadhyay, N., Chakraverty, P., Banerjee, S., Sarkar, Z. and Ghosh, S., 2021. Indian sign language classification (ISL) using machine learning. American Journal of Electronics & Communication, 1(3), pp.17-21.

[16] Rajam, P. Subha, and G. Balakrishnan. "Design and development of tamil sign alphabets using image processing with right hand palm to aid deaf-dumb people." IETE J. Res. 59.6 (2013): 709-718

[17] P. S. Rajam and G. Balakrishnan, "Real time Indian Sign Language Recognition System to aid deaf-dumb people," 2011 IEEE 13th International Conference on Communication Technology, Jinan, China, 2011, pp. 737-742, doi: 10.1109/ICCT.2011.6157974

[18] Halder, A. and Tayade, A., 2021. Real-time vernacular sign language recognition using mediapipe and machine learning. Journal homepage: www. ijrpr. com ISSN, 2582, p.7421.

Min-Kaapiyam: A Generative AI Framework based on Tholkappiyam

Balasundaram Ramaswamy

ABSTRACT:

Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics. Generative AI Models are used to generate captions (image titles) for images by first identifying the objects (nouns) and then generating a sentence that describes the image.

However Even the most advanced of present day GPTs like DALL•E 2's language understanding has limits. It is sometimes unable to distinguish "A yellow book and a red vase" from "A red book and a yellow vase" or "A panda making latte art" from "Latte art of a panda".[It generates images of "an astronaut riding a horse" when presented with the prompt "a horse riding an astronaut".

Tholkappiyam, the Tamil grammar behind the highly visual Sangam poems, is hierarchical model which can be applied to the subject of a poem (Uri-Porul) into poetic texts which invoke visual imagery in the reader's mind. Thol-Kappiyam's multi-layers like Ezuthu, Col, Verrumai, Meippaadu, Ani and Thinnai can be applied to the CLIP Image Pre-Training Algorithm to significant improve the visual quality. The integration of Thol-Kappiyam with Generative-AI to generate visual images for the Sangam poetry can act as practical technology tool to investigate the structures of Thol-Kappiyam Model as well. The generated visuals will help readers understand and appreciate the Sangam poems.

Balasundaram Ramaswamy
Entrepreneur Infony, Email: balasundaram@yahoo.com

1. METHODOLOGY:

1.1 Word2Vector – Vector Representation of Words

Word2Vector is one of the Major breakthroughs in the history of computational linguistics and NLP. Word2Vec [1] provide an efficient and automated methodology to acquire word-meaning representations. It was able to provide probabilities of next-word occurrences for a given word (such as to predict what would be most common word which will be coming next when a word is typed as in email and chat applications). It was also able very effective in other applications of Natural language processing such as text-summarization, Text Search etc. Word2Vec model also provided the basis for the Transformer Models [2] which were mappings between two related word sequences such as the mappings between the question and answer text paragraphs in a chat context.

1.2 Image Caption Generation Models

The Transformer Models were then applied to other applications such as creating the titles or captions for a given image using the Machine Learning Models of Image-Text mappings of previously given data samples [3]. The Model was to generate the text captions for a new novel given image. These models were the first models which were providing the inter-links between texts and images.

1.3 Generative Adversarial Networks for Image Generation and Picture Theory of Language

GAN or Generative Adversarial Networks were the next generation of AI models which actually reversed the functionality of image captioning models. Instead of generating text captions from images, the GAN [4] generated the images from the given text-captions.

Even though the Deep-Learning AI methodologies do not leverage any grammatically models and are fully based on statistical quantitative models, the GANs can be taken as a validation for the picture theory of language [5] put forward by the mathematician Ludwig Wittgenstein.

1.4 GPT and Generative AI Frameworks

The Generative Functionality inherent in the GAN models was leveraged to generate other type of output such as chat-answers for questions in Question-Answering systems. The Novelty of the GAN models where that these systems were synthesizing new output instead of just retrieval of pre-existing data or content in the training datasets.

Generative Pre-Trained Transformers (GPT) [7] was a major advancement within the family of generative AI models. OPENAI, the leading private research organization's DALL-E's [8] generative capabilities were accepted as a major break-through and key milestones for AI systems in general.

For example, the astronomer riding a horse image generated by GPT-2 model given below became very viral and caught the general audience's attention apart from the research community. The field of Generative AI had emerged and started to grow rapidly after the release of DALL-E2.



Fig 1: Astronaut riding a horse

DALL-E which is one of the most advanced models in the current family of Generative AI algorithms, use two key algorithms as follows:

CLIP (Contrastive Language-Image Pre-training) is responsible for recognizing text and creating a sketch of the future image;

GLIDE is responsible for converting the sketch into a final low-resolution image;

However Even the most advanced of present day GPTs like DALL•E 2's language understanding has limits. It is sometimes unable to distinguish "A yellow book and a red vase" from "A red book and a yellow vase" or "A panda making latte art" from "Latte art of a panda".[It generates images of "an astronaut riding a horse" when presented with the prompt "a horse riding an astronaut".

The linguistic and philosophical implications of the inter-relationship between the words and pictures which are generated by the Generative AI model have only been started now. The theoretical framework and insights which are offered by Ludwig Wittgenstein's language-picture model can be effectively leveraged to further advance the generative capabilities of the generative AI models.

1.5 Tholkappiyam: Poems as Painting-using-Words:

Tholkappiyam which is one of the oldest grammatical treatises in the world also uses a picture semantic model which can be termed as "Painting-using-words". In this section, the picture-semantic model of Tholkappiyam is first detailed. This interpretation of Tholkappiyam can be leveraged for creating AI-models which can augment that current generative-AI's gaps in understanding the link between sentences and images.

Before Tholkappiyam period, the poems and songs were folk-art based on mythological elements. As a great cultural leap, the aim of Tholkappiyam was to create a structured framework for composing poems whose themes and ideas where a quantum leaps over folk-art. To achieve this great theoretical leap, Tholkappiyam imagines a poem as a painting done in the mind using just word as visualization tools. Tholkappiyam prescribes the poet to first describe the background (place and time), the central characters and the decorative aspects as foreground (karu-porul). The theme (Uriporul) of the poem according to Tholkappiyam is the emergent sentiment.

The next brilliant innovation in Tholkappiyam is the usage of thinnai as key construct for elegantly describing the foreground and background of a poem. For example, if the poet sets the poem in a kurinji setting, the readers can immediately intuit the historical, geographical, ecological, psychological and sociological aspects of the poem with just a few words. The Agathinnai and Purathinnai framework also provides a genre-like construct which sets the readers expectation of what the poem would be about.



உள்ளார்கொல்லோ-தோழி!-கிள்ளை
வளை வாய்க் கொண்ட வேப்ப ஒண் பழம்
புது நாண் நுழைப்பான் நுதி மாண் வள் உகிர்ப்
பொலங் கல ஒரு காக ஏய்க்கும்
நிலம் கரி கள்ளிஅம் காடு இறந்தோரே?

THALAIVAN HAS GONE IN SEARCH OF WEALTH AWAY FROM HIS FAMILY. BEING AWAY FROM THALIAVAN MAKES THALAIVI SUFFER. THALAIVI'S FRIEND SEEING HER SUFFERING TELLS HER THAT .....

Source: https://karkanirka.org/2010/08/08/

Fig 2: A Visual Comic Image created from a Sangam text

1.6 Hierarchical Emergence in Literary Theory:

Tholkappiyam leverages and applies the hierarchical emergence framework to create a literary framework where the meaning of a poem (Uri-Porul) emerges out. Even though the book is structured as three chapters or athikaram as Ezuthu, Col and Porul athikarams, the intermediate emergence levels are multi-layered. The following is brief about the intermediate emergent layers in the literary theory.

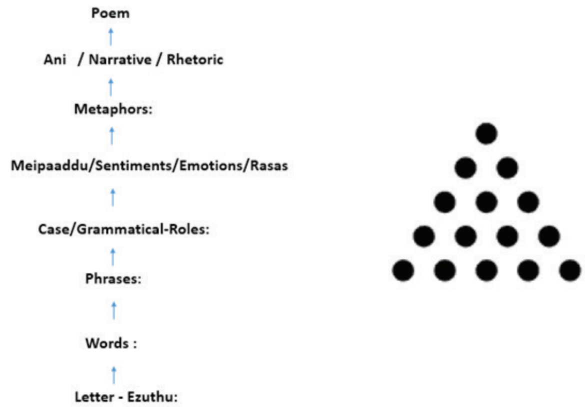


Fig 3: The Hierarchical Emergent Structure of a Poem (some levels omitted for brevity)

a. Letter - Ezuthu:

Tholkappiyam starts with the bottom most possible layer, the letter or alphabet. The alphabets are pictograms with a special meaning as sound unlike earlier writing systems where each pictogram would represent an object like bird, eye etc. It is very important to note that Tamil word for alphabet is "Ezu-thu" where "Ezu" itself indicates "emergent". Just like the word itself denotes the concept of alphabet elegantly, the grammar next defines that words are three types. The vowels and consonants are building blocks of the third type of letter uyirmei. Against unlike English or any other language, the metaphor of Uyir and Mei indicates the union of opposites and emergence very clearly.

b. Words:

Words are the next level of emergent layer. It is important to note that we can layers to forms phrases which are again not present in many languages (eg. Adukku thodar, Irattai-Kilavi) etc. Another interesting aspect of Tamil is that we can add many verbs together to form compound verbs (ottuthal) as Tamil is an agglunative language.

c. Phrases:

In Tholkappiyam, we have only two types of phrases - the Verb phrase and Noun phrase unlike later grammatical traditions. Sentences are union of these two opposites representing actions/events and objects/agents. According to Tholkappiyam, Nouns are emergent of verbs like Vedan is the one who does vettai repeatedly.

d. Case-Roles:

Next to Phrases is the grammatical or case roles which describes "who did what, when and where" in a sentence. In Tholkappiam, we need to note at the word for the agent who is responsible for action described in the sentence is called "Ezu-vai". Another important

point we need to note about Case-Roles is they are used to describe a "scene" at sentence and micro-event level.

e. Meipaaddu:

The next emergent level is emotional responses of the people involved in events. This level is not covered as part of linguistics and grammar in any other grammatical traditions. This unique layer is not what Tholkappiyam uses to link the sentences and other lower linguistics units with the thematic content, sentiment and meaning of a poem.

f. Metaphors:

Even in many modern grammatical traditions, only nouns, verbs, adverbs, adjectives etc. which are called "parts of speech" have been the core elements of grammar. Only in recent grammatical traditions like Cognitive Linguistics etc., Metaphors, Irony etc. called "Figures of Speech" have been included. Tholkappiyam goes one step further by closely leveraging its previous level in the hierarchy, the emotional layer with metaphor layer. Metaphors united two completely different and opposite objects with each other through some qualitative aspects (eg. MaaNVizhi)

g. Ani:

The final and most important layer is the rhetoric layer which is the Arrangement of Metaphors. The series of the metaphors can either praise or condemn a subject/person in the poem as the meaning of the individual metaphors combine and contribute to the creation of meaning of the song and the intention of the poet.

h. Ainthinnai:

We should also view that thinnai, the fivefold division of land, should also be viewed as emergent. Each of the five-fold thinnais like Kurinji stands as a emergent abstraction for the land, environment, flowers, ecology, people etc.

1.7 Min-Kappiyam:

Software Framework based on Tholkappiyam Grammar for Generative AI

These sections of Tholkappiyam can be modelled as individual modules of a grammatical software framework which can work coherently with a generative AI systems.

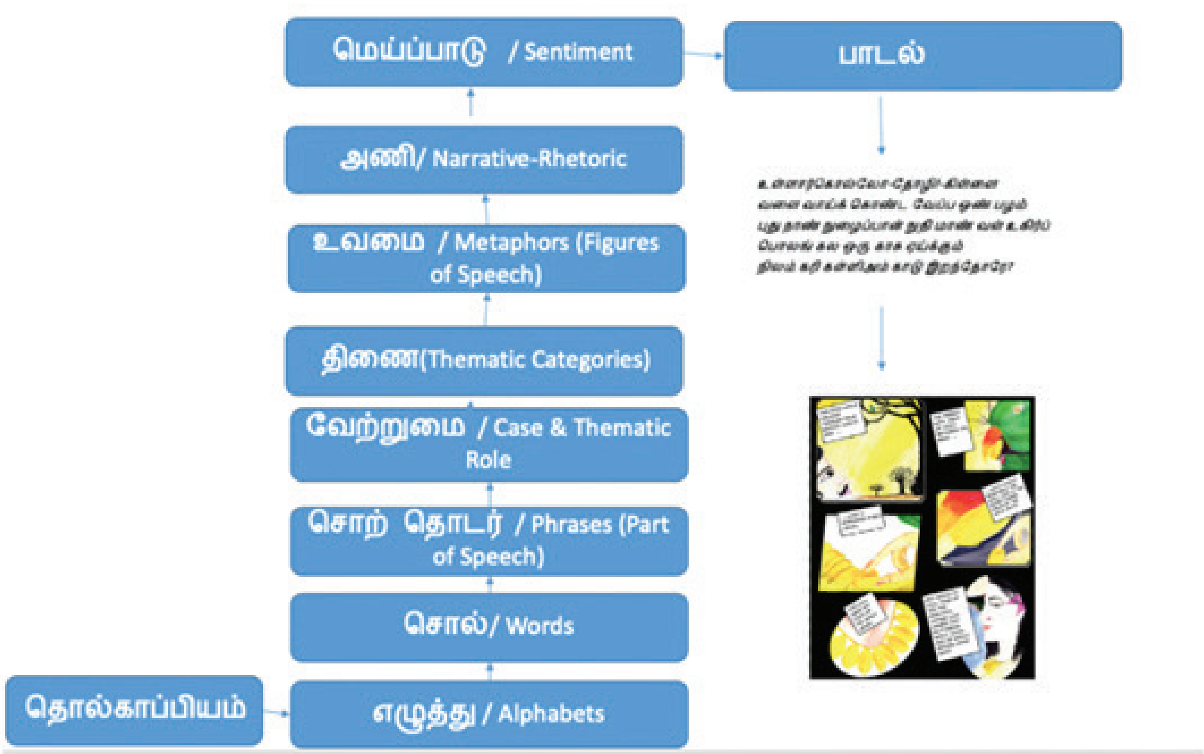


Fig 4: Software Architecture Model for Min-Kaapiyam (digital version of Tholkaapiyam)

When the Min-Kaapiyam software module is coupled with standard generative AI frameworks, we would be able to generate visual images of the sangam texts similar to how a human artiste would be able to visualize.



Fig 5: Sangam poem visuals as output of Min-Kaapiyam

1.8 Further Enhancements to the Min-Kaapiyam Model

In order for the visual images of the Min-Kaapiyam to be very effective, we might have to use completely native Tamil and Indian culture images as the pre-trained datasets. The GPT customization technique such as LORA technique can be used for this purpose. Alternatively, the training set can be also completely based on culture native content.

1.9 Future Research Perspective

The Digital Model of the Tholkappiyam can validate the unique grammatical framework of the Tamil tradition especially analysing in conjunction with Ludwig Wittgenstein philosophical framework and interconnection of Tholkappiyam with other art-forms of Tamil. This research also opens up interesting links between the "Kaatchi-Aiyam-Thelivu" Framework of Tamil Philosophical Tradition and Tholkappiyam.

REFERENCES

1. 1. Efficient Estimation of Word Representations in Vector Space Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean https://arxiv.org/abs/1301.3781

2. Attention Is All You Need Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin https://arxiv.org/abs/1706.03762

3. Bottom-up and top-down attention for image captioning and vqa. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. arXiv preprint arXiv:1707.07998 (2017).

4. Generative Adversarial Networks Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio https://arxiv.org/abs/1406.2661

5. Image captioning model using attention and object features to mimic human image understanding Muhammad Abdelhadie Al-Malla, Assef Jafar & Nada Ghneim https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00571-w

6. Picture Theory of Language https://en.wikipedia.org/wiki/Picture_theory_of_language

7. Improving Language Understanding by Generative Pre-Training Alec Radford OpenAI alec@openai.com Karthik Narasimhan OpenAI karthikn@openai.com Tim Salimans OpenAI tim@openai.com Ilya Sutskever OpenAI ilyasu@openai.com

   https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

8. Improving Image Generation with Better Captions James Betker*† jbetker@openai.com Gabriel Goh*† ggoh@openai.com Li Jing*† lijing@openai.com Tim Brooks† Jianfeng Wang‡ Linjie Li‡ Long Ouyang† Juntang Zhuang† Joyce Lee† Yufei Guo† Wesam Manassra† Prafulla Dhariwal† Casey Chu† Yunxin Jiao† Aditya Ramesh*† aramesh@openai.com

   https://cdn.openai.com/papers/dall-e-3.pdf

Review and Comparison of Tamil Text-to-Speech Systems

A Dinesh Babu

ABSTRACT

This article presents a comprehensive review and comparison of existing Tamil Text-to-Speech (TTS) systems in the hopes that it would help bridge the digital divide and make the language more accessible. We take a systematic look at the features, capabilities, restrictions, and target audience of several Tamil TTS systems, focusing on things like voice quality, naturalness, speaker customisation, language compatibility, and accessibility features. Empirical evaluations use metrics like intelligibility, prosody, and emotional expressiveness to compare systems. There are a number of Tamil TTS systems highlighted in the study, and each has advantages and disadvantages. The benefits include features like high-fidelity voices and dialectal support, while the drawbacks include things like gender bias and limited customization options. Empirical comparisons provide criteria for expressiveness and voice quality to direct future development endeavours. Using metrics like word error rate, mean opinion score, and emotional recognition accuracy, we quantitatively compare the performance of different systems. These comparisons might help users choose the best TTS for their specific needs. Using metrics like word error rate, mean opinion score, and emotional recognition accuracy, we quantitatively compare the performance of different systems. These comparisons might help users choose the best TTS for their specific needs.

A Dinesh Babu

Assistant Professor (Senior Grade), Department of ECE, SRM Institute of Science and Technology, Vadapalani, Chennai. Email: dineshba@srmist.edu.in

I. INTRODUCTION:

Early Tamil text-to-speech (TTS) systems were slow and reliant on rules to produce robotic declarations when they first appeared in the late 1990s. However, a thriving symphony of development has resounded during the last 20 years, emanating from those humble origins. The transformation was chronicled in research publications, which offered insight into the early shortcomings of clumsy algorithms and the lack of strong training data. A rainbow of Tamil TTS offers today illustrates how far the sector has come. Vocal fidelity, dialectal accuracy, emotional expressiveness, and platform integration are some of the areas where commercial ventures like SapLabs and Aravind Speech Labs compete with open-source initiatives like Kalidas and government-backed projects like the IIT Madras TTS. But dissonant sounds remain even as the richness of this choir of synthetic voices increases. Tamil TTS has obstacles that hinder its full potential, including as gender bias, restricted expressiveness, and lack of standardized resources. Scientists are resolute, mounting a counterattack by painstakingly building more extensive and varied datasets, constructing complex deep learning systems, and encouraging an open-source community mindset. A Tamil TTS that embraces complexity, diversity, and emotional eloquence—beyond basic articulation—may be on the horizon, as the future hints to. Despite their initial disagreement, this harmonious choir of voices has the ability to unite people across borders, strengthen educational and communication networks, and, in the end, delight listeners throughout the globe with the lively spirit of Tamil music.

II. REVIEW OF EXISTING TAMIL TTS SYSTEMS:

This research introduces a groundbreaking approach to combat the robotic and inaccurate voices of existing Tamil Text-to-Speech (TTS) systems. It uses Hidden Markov Models (HMMs)[1] and a carefully curated blend of speech features to analyze Tamil pronunciation and prosody. The system analyzes input text for individual phonemes, diphone combinations, and prosodic factors like stress and intonation, mapping each element to corresponding acoustic vectors from Tamil speech recordings. The system's core lies in its

specialized HMM training, which trains multiple HMMs focusing on specific aspects of speech generation. This targeted approach allows the system to capture the subtle nuances of different sounds and their interactions within a sentence, leading to smoother transitions and more natural-sounding speech.

During the synthesis stage, the system intelligently selects the most appropriate HMMs based on the features extracted from the input text. This intricate dance of feature extraction, targeted HMM training, and dynamic selection results in a significant improvement in speech quality, showcasing a 6% increase compared to traditional methods. The modular structure offers remarkable flexibility and adaptability, allowing for further refinement by incorporating additional features or expanding the system's capabilities to handle diverse regional dialects and expressive styles. However, the study has primarily focused on specific text genres and may benefit from broader evaluation across diverse content. In conclusion, this research marks a significant leap forward in the quest for high-quality and accessible Tamil TTS.

The global growth of Information and Communication technologies has led to a greater focus on speech technologies, particularly for visually impaired and vocally challenged individuals. In Tamil Nadu, India, there is a great demand for speech technology-enabled devices to facilitate access to technological and communicative facilities. However, the quality of speech in Tamil requires improved quality. This paper investigates available prosodic models [2] and details on the prosodic parameters that contribute towards improving the quality of speech synthesis stems. The study streamlines the method to develop an intonation model, which is one of the important prosodic parameters to accomplish the quality in terms of naturalness of the produced speech.

The text-to-speech synthesis system takes a series of words as input and generates speech as output. There are three major methods available to produce speech: formant synthesis, articulatory synthesis, and concatenative synthesis. The study proposes an intonation model using neural networks for a Tamil Text-to-Speech synthesis system, which uses positional, contextual, phonological, and articulatory features to train the system. The main advantage of the proposed model is the eradication of production constraints for feature extraction. The quality of synthesized speech with FO values prediction using FFNN is better than the Fujisaki model.

III. COMPARISON OF TAMIL TEXT-TO-SPEECH SYSTEMS

A. Google Text-to-Speech (GTTS):

GCTS can synthesize natural-sounding speech in over 50 languages and a variety of voices, from classic to emotional to character-specific. It also supports speech effects like background noise, pitch shifting, and speaking rate adjustments. GCTS offers several advantages, including high-quality audio, a wide range of languages and voices, flexible integration with other Google Cloud services, and scalability to handle large workloads. GCTS can be used for a variety of applications, such as creating voiceovers for videos, adding narration to e-learning modules, building interactive voice assistants, and personalizing customer experiences. GCTS uses a pay-as-you-go pricing model, with costs based on the number of characters synthesized and the chosen voices. You can also take advantage of a free tier for low-volume usage. GCTS is easy to use, with various client libraries and SDKs available for different programming languages. You can also use the web interface to try it out without any coding.

Technical details: Server-based neural network model, LSTM architecture, multiple voice options, adjustable rates.

Strengths: High speech quality, user-friendly interface, integration with Google Cloud Platform.

Weaknesses: Limited customization options, potential cost limitations for advanced features.

B. Microsoft Azure Text-to-Speech (MA TTS):

Microsoft Azure Text-to-Speech (MA TTS) is a cloud-based service that transforms text into lifelike speech, offering over 270 voices in 119 languages and customization options. It's integrated with Azure services for building intelligent voice experiences, and is used for voiceovers, audio books, voice assistants, assistive technologies, customer service, and content personalization.

Technical details: Deep neural network architecture, hybrid statistical and deep learning approach, speaker adaptation functionalities.

Strengths: Wide range of voices, customizable prosody, high-quality output for specific domains.

Weaknesses: Complex pricing structure, high compute demands for advanced features.

C. Resemble.ai:

Resemble.ai, your cloud-based AI voice playground, lets you clone or transform voices, speak in 60+ languages, and even sniff out deepfakes. From crafting personalized audiobooks to building multilingual

chatbots, Resemble.ai empowers you to create, control, and protect voices for limitless possibilities.

Technical details: Deep learning and VAE based architectures, expressive and realistic speech synthesis, speaker cloning capabilities.

Strengths: Highly natural and expressive output, suitable for creative applications.

Weaknesses: Expensive for commercial use, advanced customization requires technical expertise.

D. Vokaturi:

Vokaturi, your personal voice alchemist, lets you morph and mix audio like magic. Its AI engine seamlessly blends voices, creates realistic deepfakes, and even extracts emotions from sound. Craft unique narrations, build expressive chatbots, or simply have fun experimenting – Vokaturi puts the power of voice manipulation in your hands.

Technical details: HMM-based synthesis, flexible and customizable architecture, multiple dialects support.

Strengths: Technical knowledge empowers customization, suitable for research and development purposes.

Weaknesses: Less natural speech quality compared to other systems may require deeper technical understanding.

Table I. Comparison of various TTS systems

Feature	GTTS[4]	MA TTS[3]	Resemble.ai[6]	Vokaturi [5]
Architecture	Neural network	Deep neural network	Deep learning & VAE	HMM
Speech quality	High	High	Highly natural & expressive	Less natural
Customization	Limited	High	Advanced	Flexible
Cost	Free (basic)	Pay-per-use	Expensive	Free (open-source)
Target users	General users, developers	Researchers, developers, advanced users	Content creators, media professionals	Developers, researchers

IV. ADVANCEMENTS IN TAMIL LANGUAGE TECHNOLOGY

The field of Tamil language technology has witnessed significant advancements in recent years, thanks to the continuous efforts of researchers, developers, and language enthusiasts. These advancements have contributed to the improvement of Tamil text-to-speech systems and other language-related technologies.

One notable advancement is the integration of deep learning techniques into text-to-speech systems. Deep learning models, such as recurrent neural networks and convolutional neural networks, have shown promising results in enhancing the naturalness and intelligibility of generated speech. By leveraging these techniques, developers have been able to create more realistic and human-like Tamil text-to-speech systems.

Another area of advancement is the development of domain-specific text-to-speech systems. These systems are designed to cater to specific domains or industries, such as healthcare, finance, or legal. By fine-tuning the models with domain-specific data, developers can improve the accuracy and quality of speech output in specialized contexts.

Furthermore, advancements in data collection and processing have led to the creation of larger and more diverse datasets for training text-to-speech models. These datasets include a wide range of Tamil texts, ensuring better coverage of various linguistic aspects. With access to more extensive and representative datasets, developers can create more robust and accurate text-to-speech systems.

V. CHALLENGES IN DEVELOPING TAMIL TEXT-TO-SPEECH SYSTEMS

While advancements in Tamil language technology are commendable, there are still several challenges that developers face when creating Tamil text-to-speech systems. These challenges can hinder progress and limit the effectiveness of the systems. Some of the key challenges include:

Limited resources and funding: Developing high-quality text-to-speech systems requires substantial resources and funding. However, the availability of such resources is often limited, hindering the development of advanced systems.

Linguistic complexities: Tamil, like any other language, poses certain linguistic complexities that make it challenging to develop accurate text-to-speech systems. Tamil's rich morphology, complex grammar, and phonetic variations require careful modelling and handling.

Contextual understanding: Text-to-speech systems need to understand the context in which the text is being spoken to ensure appropriate pronunciation and intonation. Incorporating contextual understanding into the systems remains a challenge, particularly for complex or ambiguous sentences.

Voice variety and customization: Users often have diverse preferences when it comes to voices. Incorporating voice variety and customization options in text-to-speech systems can be challenging due to limited resources and technical constraints.

Overcoming these challenges requires collaborative efforts from researchers, developers, and stakeholders, along with increased support and investment in Tamil language technology.

VI. PROMOTING AWARENESS AND ADOPTION OF TAMIL TEXT-TO-SPEECH SYSTEMS

Creating awareness and promoting the adoption of Tamil text-to-speech systems is essential for maximizing their impact and ensuring widespread accessibility. Here are some strategies to promote awareness and encourage the use of these systems:

Education and outreach programs: Organize workshops, seminars, and training sessions to educate individuals about the benefits and applications of Tamil text-to-speech systems. These programs can target schools, universities, libraries, and organizations working with visually impaired individuals.

Collaboration with content creators: Collaborate with content creators, such as publishers, e-learning platforms, and digital media companies, to integrate Tamil text-to-speech systems into their platforms. This collaboration can help raise awareness among content creators and encourage the adoption of these systems.

User feedback and improvement: Encourage users to provide feedback on their experiences with Tamil text-to-speech systems. This feedback can help developers identify areas for improvement and enhance the overall user experience.

Policy and advocacy: Advocate for policies that promote the integration of Tamil text-to-speech systems in public and private digital platforms. Engage with policymakers, language organizations, and advocacy groups to ensure the inclusion of Tamil language technology in the digital landscape.

By implementing these strategies, we can promote awareness, drive adoption, and foster the growth of Tamil text-to-speech systems, ultimately enhancing language accessibility for Tamil speakers.

VII. FUTURE PROSPECTS AND DEVELOPMENTS IN TAMIL LANGUAGE TECHNOLOGY

The future of Tamil language technology looks promising, with several exciting developments on the horizon. Some potential areas of growth and advancement include:

Voice customization: Future developments may focus on enhancing voice customization capabilities, allowing users to create personalized voices based on their preferences and requirements.

Emotion and expression: Integrating emotion and expression into text-to-speech systems can add a new dimension to speech synthesis. Future systems may incorporate emotional cues and variations to make the generated speech more engaging and expressive.

Real-time applications: Real-time text-to-speech systems can enable live translation, transcription, and voice assistance in various domains, including education, customer support, and healthcare. Future developments may explore real-time applications to facilitate seamless communication and accessibility.

Multilingual support: Expanding the capabilities of Tamil text-to-speech systems to support multiple languages can enhance their versatility and usefulness. Future developments may aim to integrate multilingual capabilities, enabling users to switch between different languages seamlessly.

The future of Tamil language technology relies on constant innovation, collaboration, and support from various stakeholders. By embracing these developments, we can unlock the full potential of Tamil text-to-speech systems and revolutionize language accessibility for Tamil speakers.

VIII. CONCLUSION

In conclusion, Tamil text-to-speech systems play a vital role in promoting language accessibility, inclusivity, and cultural heritage. By reviewing and comparing various systems, we gain insights into their strengths,

weaknesses, and potential areas of improvement. The advancements in Tamil language technology, coupled with the challenges faced in developing text-to-speech systems, highlight the need for continuous collaboration and support. Promoting awareness and adoption of Tamil text-to-speech systems is crucial for maximizing their impact. By educating individuals, collaborating with content creators, and advocating for policies, we can ensure the integration of these systems into the digital landscape. The future of Tamil language technology holds immense potential for voice customization, real-time applications, and multilingual support. By embracing these prospects and fostering innovation, we can create a future where language accessibility knows no boundaries. Let's continue to push the boundaries of Tamil language technology and unlock a world of possibilities for Tamil speakers worldwide.

REFERENCES

[1] Jayakumari, J., & Jalin, A. F. (2019, June). An Improved Text to Speech Technique for Tamil Language Using Hidden Markov Model. 2019 7th International Conference on Smart Computing & Communications (ICSCC). https://doi.org/10.1109/icscc.2019.8843683

[2] Rajeswari, K. C., & UmaMaheswari, P. (2014, December). A novel intonation model to improve the quality of tamil text-to-speech synthesis system. In 2014 Sixth International Conference on Advanced Computing (ICoAC) (pp. 722-727). IEEE. https://doi.org/10.1109/ICoAC.2014.7229737

[3] https://azure.microsoft.com/en-us/products/ai-services/speech-translation

[4] https://gtts.readthedocs.io/en/latest/

[5] https://vokaturi.com/

[6] https://www.resemble.ai/

[7] H. A. Patil et al., "A syllable-based framework for unit selection synthesis in 13 Indian languages," 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurgaon, India, 2013, pp. 1-8, doi: 10.1109/ICSDA.2013.6709851.

[8] D. S. S. De Zoysa, J. M. Sampath, E. M. P. De Seram, D. M. I. D. Dissanayake, L. Wijerathna and S. Thelijjagoda, "Project Bhashitha - Mobile Based Optical Character Recognition and Text-to-Speech System," 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, Sri Lanka, 2018, pp. 1-5, doi: 10.1109/ICCSE.2018.8468858.

[9] A. F. Jalin and J. Jayakumari, "Text to speech synthesis system for tamil using HMM," 2017 IEEE International Conference on Circuits and Systems (ICCS), Thiruvananthapuram, India, 2017, pp. 447-451, doi: 10.1109/ICCS1.2017.8326040.

[10] M. Karthikadevi and K. G. Srinivasagan, "The development of syllable based text to speech system for Tamil language," 2014 International Conference on Recent Trends in Information Technology, Chennai, India, 2014, pp. 1-6, doi: 10.1109/ICRTIT.2014.6996126.

TAMIL INFORMATION RETRIEVAL AND TEXT MINING

தமிழ்க் கணினி ஆய்வுகள்: சவால்களும் எதிர்காலமும்

வாசு அரங்கநாதன்

ஆய்வுச்சுருக்கம்

தமிழ்க் கணினி ஆய்வு என்பதைத் தமிழின் பல ஆய்வுகளோடு ஒப்பிட்டு எவ்வகையிலெல்லாம் கணினி கொண்டு அவ்வாய்வுகளுக்குப் புதிய கண்ணோட்டத்தைக் கொடுக்கலாம் என்று சிந்திக்க வேண்டிய நிலை. தமிழியல் என்பதை இக்காலகட்டத்தில் தமிழ்க் கணினியியல் எனக் கூற வேண்டியிருக்கிறது. ஏனெனில் பெரும்பாலான தமிழ் ஆய்வுகளைக் கணினி கொண்டு ஆய்வு செய்ய வேண்டிய கட்டாய நிலையில் இருக்கிறோம். இலக்கிய ஆய்வு, கல்வெட்டியல், அகழாய்வு, தமிழ்க் கற்றல் மற்றும் கற்பித்தல், மொழியியல் போன்ற ஆய்வுகளைக் கணினி இன்றி செய்ய இயலாத நிலைக்குத் தள்ளப்பட்டிருக்கிறோம். மொழியியல் ஆய்வைக் கணினி மொழியியல் ஆய்வு (computational linguistics) எனவும் கணினி நுண்திறன் ஆய்வு (artificial intelligence/natural language processing) எனவும் இருவேறாகப் பிரித்திருக்கின்றனர். முன்னது கணினி கொண்டு வட்டார வழக்கு ஆய்வு, சமூக மொழியியல் ஆய்வு போன்ற ஆய்வுகளை மேம்படுத்துவது. பின்னது கணினிக்கு மொழித் திறனைக் கொடுத்து மனிதன் செய்யும் மொழிப் பயன்பாடுகளைச் கணினியைச் செய்ய வைப்பது. இப்பிரிவில் உரையைப் புரிந்துகொள்ளும் திறன், உரையிலிருந்து பேச்சு, பேச்சிலிருந்து உரை, மொழிபெயர்ப்பு போன்ற மொழித் திறன்களைக் கணினிக்குக் கொடுப்பது அடங்கும். ஆனால் இலக்கிய ஆய்வு, கல்வெட்டியல் ஆய்வு, அகழாய்வு, தமிழ்க் கற்றல் மற்றும் கற்பித்தல் ஆகிய ஆய்வுகளில் இத்துறைகளை எப்படியெல்லாம் கணினி கொண்டு மேம்படுத்தலாம் என்று காணும் சூழல். இத்துறைகளைக் கணினி கொண்டு ஆய்வு செய்யும் புதிய அணுகுமுறைகளைக் கடைப்பிடிக்காவிட்டால் நமது ஆய்வு மேம்படவில்லை எனவே கூற வேண்டிய சூழலில் இருக்கிறோம். எடுத்துக்காட்டாகக் கணினி கொண்டு இலக்கியத் தரவுகளை ஆய்வு செய்யும் முயற்சியின் வழித் தமிழ் மொழியில் கொண்டிரு, விடு, கொள் போன்ற உருபுகள் எவ்வாறு இடைக்காலத்தில் உருவாகின என்பது, "வரல் ஆகும்" என்றிருந்த வழக்கு எப்படி "வரலாம்" என ஆயிற்று, "அகன்றிசின்" என்பது எப்படி "ஆன்றிசின்" என மாறியது போன்ற பல செய்திகளை நுணுக்கமாக அறியும் முறையை அறியலாம். சங்ககாலத்தில் இல்லாத இவ்வுருபுகள் இடைக்காலத்தில் எவ்வாறு ஏற்பட்டன என்னும் செய்தியைப் படிப்படியாகத் தமிழ் மொழியின் வரலாற்றடிப்படையில் அறிய வேண்டுமெனில் கணினி வழி இலக்கியத் தரவுகளை ஆய்வு செய்யும்போதுதான் அறிய முடியும். (காண்க. ரெங்கநாதன் 2011, 2018, 2023). கணினி கவிதை எழுதுவது, கதைகளை எழுதுவது போன்ற திறனை இக்காலக் கணினி நுண்திறன் ஆய்வில் செய்திருந்தாலும் இவ்வகை முயற்சிகளில் மனிதனின் இலக்கிய உருவாக்கத் திறனுக்கு இயந்திரம் ஈடுகொடுக்க இயலாத நிலையில்தான் இருக்கிறது எனலாம்.

வாசு அரங்கநாதன்
பென்சில்வேனியாப் பல்கலைக் கழகம், அமெரிக்கா.
Email: vasur@sas.upenn.edu

முன்னுரை:

இக்கட்டுரையில் கடந்த இருபது ஆண்டுகளுக்கும் மேலாகப் பென்சில்வேனியாப் பல்கலைக்கழகத்தில் நாங்கள் ஈடுபட்ட கணினித் தமிழ் ஆய்வு பற்றி விளக்கி எதிர்காலத்தில் முறையாக எது போன்ற கணினித் தமிழ் ஆய்வில் நாம் ஈடுபட வேண்டும் என்பதை இங்கு விளக்க முற்படுகிறோம். இக்கட்டுரையில் முதலாவதாக தமிழ்க் கற்றல் மற்றும் கற்பித்தல் துறையில் கணினியை எவ்வாறெல்லாம் பயன்படுத்தி வருகிறோம் என்பது பற்றிய விரிவான விளக்கத்தை அளித்து அதற்கான இணையப் பக்கங்களையும் விளக்க முற்படுகிறோம். இரண்டாவதாகத் தமிழ் இலக்கிய ஆய்வுக்கென நாங்கள் ஏற்படுத்தியிருக்கும் இணையப்பக்கங்களை விளக்கி அவற்றின் அடுத்தக் கட்ட ஆய்வுகள் பற்றியும் விளக்க முற்படுகிறோம். மூன்றாவதாகக் கணினி நுண்திறன் ஆய்வு என்னும் பிரிவின் கீழ் மனித இயந்திரம் உருவாக்கும் எங்களது ஆய்வை விளக்கி மனித இயந்திரத்தோடு தமிழில் உரையாடும் வழிமுறைகளையும் விளக்குகிறோம். [1]

1. கணினி வழித் தமிழ் கற்றலும் கற்பித்தலும்

வம்சாவளியினரின் பயன்பாட்டுக்கேற்ப கணினி வழித் தமிழ்க் கற்றலும் கற்பித்தலுக்குமான இணையப் பக்கம் மிகவும் முக்கியம் இக்காலக்கட்டத்தில். இந்நோக்கத்தில் எண்ணற்றப் பக்கங்கள் இருக்கின்றன. இருப்பினும் இரண்டாம் மொழியாகத் தமிழ் மொழியைக் கற்கும் மற்றும் கற்பிக்கும் நோக்கில் இதற்கான இணையப் பக்கத்தை ஏற்படுத்துவதில் நாம் மிகவும் கவனமாக இருக்க வேண்டும். இவ்வழியில் அமைக்கப்பட்ட பின்வரும் இணையத் தளங்களைப் பற்றி இப்பகுதியில் காண்போம்.

http://tamilverb.com



http://learn.tamilnlp.com



இவ்விரண்டு இணையப் பக்கங்களும் பென்சில்வேனியாப் பல்கலைக்கழகத்தின் தெற்காசியத் துறையில் செய்யப்பட்ட ஆய்வின் அடிப்படையில் உருவாக்கப்பட்டன. முதல் பக்கம் மொழிச் சூழலும் மொழிக்கற்றலும் என்ற கருத்தின் அடிப்படையில் உருவாக்கப்பட்டது. இரண்டாவது பக்கம் பேராசிரியர் ஷிப்மன் மற்றும் வாசு அரங்கநாதன் அவர்கள் ஈடுபட்ட தமிழ் வினைகள் பற்றிய ஆய்வின் அடிப்படையில் உருவாக்கப்பட்டது.

1.1. மொழிச்சூழலும் மொழிக் கற்றலும்

மொழிச்சூழல் (language in context) என்பதைப் பல கோணங்களில் விளக்கலாம். மொழியைக் கற்பது என்பது மொழியின் இலக்கணம் மற்றும் சொற்களைப் பற்றி அறிவதோடன்றி அவை பேச்சுச் சூழலில் எங்கனம் பயன்படுத்தப்படுகிறது என்பதை அறிவதும் ஆகும். "நீங்களும் வந்துவிட்டீர்களா?" என்னும் வினாவை இலக்கண அடிப்படையில் அறிந்து கொண்டால் மட்டும் இத்தொடரை எல்லா மொழிச் சூழலிலும் பயன்படுத்த முடியும் என்பதாகிவிடாது. ஒவ்வொரு சூழலுக்கும் ஒவ்வொருவிதமாக இத்தொடரைப் புரிந்துகொள்ளலாம்.[2] இந்த வகையில் "விடு" என்ற விகுதியின் பல்வேறு பயன்பாடுகளை அதைப் பயன்படுத்தும் சூழல்கள் மூலம்தான் நன்கு அறிய முடியும். என்னென்ன சூழலில் எப்படியெல்லாம் இந்த விகுதியைப் பயன்படுத்தலாம் என்பதை அறுதியிட்டும் கூறிவிடமுடியாது. மொழியை முதல் மொழியாக அறிந்தவர்கள் இவ்வகை பொருள் நுண்மைகளை பல சூழல்களில் தங்களை ஈடுபடுத்திக்கொண்டு இவ்விகுதிகள் குறித்தான பயன்பாட்டை நன்கு அறிந்துகொள்கின்றனர். இத்தகைய பொருள் நுண்மைகளை எங்கனம் இரண்டாம் மொழியாகத் தமிழ் மொழியைக் கற்கும் மாணவர்களுக்குக் கொடுக்க முடியும்? முதல் மொழியாக ஒருவர் மொழியை எங்கனம் கற்றுக்கொண்டாரோ அதே சூழலை மொழி கற்றல் முறையில் கொடுப்பதே பயன்பாட்டு அடிப்படையில் மொழிக் கற்பித்தல் என்பது.

1. இக்கட்டுரையை தமிழக அரசின் தமிழ்க் கணினி மாநாடு 2023ல் படைக்க வாய்ப்பு கொடுத்த தமிழக அரசுக்கும், தமிழ் இணையக் கல்விக்கழகத்துக்கும் எனது நன்றியைத் தெரிவித்துக்கொள்கிறேன். இருபது ஆண்டு கால தமிழ்க் கணினி ஆய்வுக்கு எனக்குப் பல ஆதரவுகளையும் தொடர்ந்து நல்கிவரும் பென்சில்வேனியாப் பல்கலைக்கழகத்தின் தெற்காசியத் துறைக்கும் எனது நன்றியை இங்கு தெரிவித்துக்கொள்கிறேன்.

2. "விடு" என்னும் விகுதியின் பயனைச் சூழலின் அடிப்படையில் காண்க: http://learn.tamilnlp.com/unit_07/section_A/lesson02.html

மாணவர்களிடையே மொழிச் சூழல்கள் பலவற்றைக் கொடுத்து மொழியைச் சரியாகப் பயன்படுத்தும் திறமையை வளர்க்கச் செய்ய வேண்டும். இதையும் இலக்கண அடிப்படையிலான மொழிக்கற்றல் முறை என விவாதிக்கலாம்! ஆனால் இலக்கணம் பற்றி நிறைய சிந்தனை இல்லாமலே மொழியைச் சூழலுக்கு ஏற்றவாறு கற்பித்தல் பயன்பாட்டு முறை மொழிக்கற்பித்தலின் சிறப்பு அம்சம் எனலாம்.

1.2. இலக்கண அடிப்படையிலான மொழிக் கற்பித்தலின் சிக்கல்கள்

மொழிக்கற்றல் துறையின் ஆரம்பகாலங்களில் கட்டமைப்பு மொழியியல் கோட்பாடுகள் (structural linguistics) பலவற்றையே பயன்படுத்திவந்தனர். இதனால் பல மொழியாசிரியர்கள் மொழியியல் கோட்பாடுகளைப் பயன்படுத்தியே மொழி கற்பிக்கும் வகையை அறிந்தனர். இதனால் மாணவர்கள் மொழியைக் கற்கும் போது பெரும்பாலும் இலக்கணத்தையே கற்றுக்கொண்டனர். இதன் பின்விளைவு மாணவர்களால் மொழியை முதல் மொழியாகப் பேசுபவர்கள் போன்ற திறமையைப் பெற முடியவில்லை! மேலும் தமிழ் மொழியைப் பொறுத்தவரையில் இன்னமும் செம்மொழி இலக்கணத்தையும் இலக்கியத்தையும் கற்பிப்பதே முதன்மையாக இருத்தல் வேண்டும் என்பதில் சிலர் கவனமாக இருக்கின்றனர். தமிழ் இலக்கியம் கற்பிப்பது என்பது வேறு தமிழ் மொழி கற்பிப்பது என்பது வேறு. இலக்கியம் மூலமாகத் தமிழ் மொழியைக் கற்பித்தல் வேண்டும் என்பதும் சிலரின் கருத்து. அதோடு மொழியாசிரியர்களுக்கு செயல்பாட்டு அடிப்படையிலும் (performance based approach), சூழல் அடிப்படையிலும் (context based approach), உரையாடல்கள் அடிப்படையிலும் (communicative approach), தேவை அடிப்படையிலும் (task based approach) மொழியை எவ்வாறு கற்பிப்பது என்பது போன்ற பயிற்சிகள் போதுமான அளவில் கிடைக்காமல் போய்விட்டது. இவ்வகைகளில் தமிழ் மொழியைக் கற்பிப்பதற்கான பயிற்சி நூற்கள் பல இல்லாதது தமிழ் மொழிக் கற்றல் முறைகளில் போதுமான மாற்றங்களை ஏற்படுத்த இயலாத நிலை! சூழல்கள் அடிப்படையில் தமிழ் கற்பதற்காக உருவாக்கப்பட்ட நூலை ரெங்கநாதன் 2011ல் காணலாம்.

1.3. புலம்பெயர்ந்த தமிழ் மாணவர்களின் தமிழ் மொழித் தேவைகள்

புலம்பெயர்ந்த தமிழ் வம்சாவளி மாணவர்களுக்கான மொழித்தேவை என்ன என்பதை ஆய்ந்தால் அவர்கள் தங்களின் மொழிக்கற்றல் சூழல் அடிப்படையில் பல்வேறு வகையான தமிழ் மொழித்திறைமையைப் பெற்றுள்ளார்கள் என்பதை அறியலாம். அமெரிக்கப் பல்கலைக்கழகங்களில் தமிழ் பயிலும் மாணவர்களைப் பொறுத்த வரையில் கொஞ்சம் கூட தமிழ் தெரியாமல் வருபவர்கள் முதல் நன்கு பேசும் திறமைப் பெற்று வரும் மாணவர்கள் வரை இம்மாணவர்களைப் பல வகையில் பிரிக்கலாம். இருப்பினும் மிகக் குறைந்த தமிழ் மாணவர்களே தங்களின் பெற்றோர்கள் மூலமும், தொலைக்காட்சி, திரைப்படம் போன்றவை மூலமும் தமிழை நன்கு புரிந்து கொண்டு பேசக் கூடிய திறமையைப் பெற்றவர்களாக இருக்கிறார்கள். பெரும்பாலான மாணவர்கள் தமிழை நன்கு புரிந்து கொள்கிறார்கள் ஆனால் அவர்களுக்குப் பேசவும் திறம்பட எழுதவும் இயலாத நிலையில்தான் இருக்கிறார்கள். இவர்களையும் மற்றவர்கள் போல் தமிழ் மொழி பேசவும் எழுதவும் செய்ய வேண்டுமெனில் பாடத்திட்டங்களை மேற்கூறிய வகையில் செயல்பாட்டுத்திறனில் அமைப்பதே சிறந்த வழியாகும். இத்தகைய புலம்பெயர்ந்த வம்சாவளி மாணவர்களுக்கு இலக்கண அடிப்படையிலோ வேறு எந்தவித வழக்கமான அடிப்படையிலோ மொழிக் கற்பித்தல் என்பது சரியான விளைவைக் கொடுக்குமா என்பது கேள்விக்குறியே!

சில தமிழ் மாணவர்கள் வீட்டில் பேசும் தமிழைக் கற்றுக் கல்லூரிக்கு வரும் போது அங்கு கற்றுக்கொடுக்கப்படும் இலக்கணம் மற்றும் இலக்கியங்கள் அவர்களுக்குப் புரியாத புதிராகவே இருக்கின்றன. இவற்றைத் தங்களின் வீட்டில் கற்றுக்கொண்ட தமிழறிவோடு இணைத்துப்பார்க்கும் வகை இவர்களுக்கு இயலாமல் போய்விடுகிறது.

தமிழ் வம்சாவளி மாணவர்களுக்குத் தமிழ் உள்ளுணர்வு (Tamil intuition) கிடைக்கச்செய்தல் வேண்டும். இந்த உள்ளுணர்வு என்பது தமிழ் மொழியின் நுண் பொருள்களைப் பற்றிய மற்றும் மொழியை முறையாகப் பயன்படுத்தும் திறமை பற்றிய அறிவாகும். இந்த உள்ளுணர்வு கிடைக்காத பட்சத்தில் அவர்கள் தமிழ் மொழியைப் பள்ளியறிவாகவே காண்பர். மற்றும் மொழிச் சூழலுக்குத் தகுந்தவாறு பேசும் திறமையைப் பெறாதவர்களாவே இவர்கள் இருப்பர். தமிழ் மொழியில் உள்ள பல சிக்கல்களை அவர்கள் அறியாமலும் பயன்படுத்தத் தெரியாமலுமே இருப்பர். உதாரணமாக "நீங்களாவது வாங்கித்தருவதாவது",[3] "வந்தோமா படித்தோமா போனோமா என்று இல்லாமல்..."[4], "வாங்க போங்கங்றீங்க..."[5] போன்ற வழக்குகளை முறையாகப் பயன்படுத்த ஒருவருக்கு இவற்றைப் பற்றிய முழுப் பயன்பாட்டு அறிவு தேவை! மொழிச்சூழல் அடிப்படையிலான இப்பயன்பாடுகளை அறிந்து கொள்வதிலும் பயன்படுத்துவதிலும் அவர்களுக்குச் சிரமமாகவே இருக்கும். இவ்வகையிலான

3. காண்க http://learn.tamilnlp.com/unit_08/section_A/lesson01.htmil
4. காண்க http:// learn.tamilnlp.com/unit_09/section_B/lesson02.html
5. காண்க http:// learn.tamilnlp.com/unit_09/section_A/lesson01.html

பயன்பாடுகளை விளக்கினால் மட்டும் அவர்களுக்கு முழுமையாகப் புரியும் வாய்ப்பு இருக்கும் என்று கூறிவிட முடியாது. மாறாக இவற்றை அவர்கள் இச்சூழல்களில் பயன்படுத்தும் வாய்ப்பு கிடைக்கும் போதுதான் இவற்றின் நுண்பொருள்களும் பயன்பாடுகளும் அவர்களின் உள்ளுணர்வில் பதிய வாய்ப்பிருக்கும். குறிப்பாக இவ்வகை பயன்பாடுகள் மொழியைப் பேசும் போது இயல்பாக அவர்களுக்கு வர வேண்டும் என்பது மிக முக்கியம். இவ்வகைத் திறனை அவர்களுக்கு பள்ளியறிவு மூலமே கொடுத்துவிட முடியுமா என்பது கேள்விக்குரியதே!

தமிழை இரண்டாம் மொழியாகக் கற்போருக்குத் தமிழைப் பயன்படுத்துவதற்கான நல்ல மொழிச்சூழல் ஈடுபாடு (language immersion) கிடைக்காத தருணத்தில் அவ்வகை ஈடுபாட்டை மொழிக்கல்வி வகுப்புகளில் கொடுக்க முனைவதே மொழிப் பயன்பாட்டு உத்தியின் பெரும்பங்காகும்.

1.4. தமிழ் மொழிக் கற்றல் மற்றும் கற்பித்தலில் வேண்டிய மாற்றங்கள்

பெரும்பாலான வம்சாவளி தமிழ் மாணவர்களுக்குத் தமிழைக் கேட்டுப் புரிந்துகொள்ளும் திறன் முழுமைப் பெற்ற நிலையில் இருப்பார்கள். அதாவது எவ்வளவு வேகமாக அவர்களிடம் தமிழில் பேசினாலும் அவர்களால் மிக நன்றாகப் புரிந்து கொள்ளமுடிகின்ற திறன் அவர்களிடம் இருக்கும். இத்தகைய மாணவர்களுக்கு மேற்கூறிய சிறப்பு மொழிப் பண்புகள், வழக்குத் தொடர் மற்றும் பல சிக்கலான அமைப்புகள் பலவற்றைப் புரிந்துகொள்கிற திறன் நன்கு இருக்கும். ஆனால் அவற்றையெல்லாம் தக்க மொழிச் சூழலில் பயன்படுத்தும் திறன் கொஞ்சம் கூட இருக்காது. புரிந்து கொள்ளும் திறன் மற்றும் செயல் திறன் என்னும் இருவகைத் திறனைப் பற்றி மொழிக்கற்பிப்போரின் மனதில் இருக்க வேண்டியிருக்கிறது. அவர்களிடம் இருக்கும் புரிந்துகொள்ளும் திறனைக் கொண்டு அவர்களிடம் இல்லாத செயல் திறனை வளர்க்க வேண்டியதே தமிழ் கற்பிப்போரின் முழுக்கவனத்தில் இருக்க வேண்டும். அவர்களின் புரிந்துகொள்ளும் திறனை முறையாகப் பயன்படுத்தாத எந்த வித பாடத்திட்டமும் அவர்கள் மத்தியில் எடுபடாமல் போய்விட வாய்ப்பிருக்கும். இந்தவகையில் பல மொழிச் சூழல்களைக் காணொளிகள் மூலம் பார்க்கச் செய்து அவ்வுரையாடல்களைக் கேட்கச் செய்து அவ்வுரையாடல்கள் போலவே வகுப்பில் சகமாணவர்களோடு செயற்கைச் சூழலை ஏற்படுத்திப் பயன்படுத்த வைப்பதே இப்பயன்பாட்டு அடிப்படையிலான தமிழ் மொழிக் கற்றல் முறையின் முக்கிய அம்சமாகும். அவர்கள் சில செயற்கைச் சூழலை ஏற்படுத்த வேண்டியிருக்கும். இருப்பினும் அத்தகையச் சூழல்களில் தங்களை ஈடுபடுத்திக்கொள்ளும் போது தங்களின் கேட்கும் திறனை பேசும் திறனாகக் கொண்டுவர வாய்ப்பிருக்கும்.

மொழிப் பயன்பாடு மற்றும் மொழிச் சூழல்கள் அடிப்படையில் அமைக்கப்பட்ட இணையப் பக்கம் http://learn.tamilnlp.com.

இவ்வகையில் மேற்கூறிய இணையப் பக்கத்தில் எழுபத்திரண்டு மொழிச் சூழல்கள் ஒளிக்காட்சிகளாகக் கொடுக்கப்பட்டிருக்கின்றன. வீடு, கடைவீதி, பேருந்து நிலையம், பழக்கடை, காய்கறிக்கடை, துணிக்கடை, சாப்பாடு, உடை போன்ற பல வகையான மொழிச் சூழல்கள் இக்காணொளிகள் வழியாகக் கொடுக்கப்பட்டுள்ளன. இவை தமிழ்மொழிப் பயன்பாடு பற்றி மட்டுமல்லாமல் தமிழ்ப்பண்பாடு மற்றும் எதார்த்தமான சூழ்நிலைகள் போன்ற விளக்கங்களையும் மனதில் கொண்டு அமைக்கப்பட்டவை! இவ்வொளிக்காட்சிகளும் அவற்றோடு கொடுக்கப்பட்டுள்ள பாடத்திட்டங்களும் பயிற்சிகளும் எளியன முதல் மிகச் சிரமமானவை என ஒரு வரையரைக்குள் கொடுத்துள்ளோம். முதல் ஒளிக்காட்சியைக் காணும் போது அது மிக எளியதான வரவேற்பு மற்றும் ஒருவருக்கொருவர் பழகிக்கொள்ளும் சூழலை விளக்குவதாக இருக்கும். பாடங்கள் தொடரத் தொடர அவை பல சிக்கலான சூழல்களை விளக்குவதாக இருக்கும். வம்சாவளி மாணவர்கள் தங்களின் மொழித் திறமையின் அடிப்படையில் இப்பாட வரிசையில் எந்தப் பாடத்திலிருந்து வேண்டுமானாலும் படிக்கத் தொடங்கலாம். இப்பாடத்திட்டங்களைப் படிக்க வேண்டும் என்று சொல்வதைவிட இவற்றோடு தங்களைப் பழக்கப்படுத்திக்கொள்ளலாம் என்று சொல்வதே மிகப் பொருந்தும். இவ்வகையில் ஒவ்வொரு உரையாடலிலும் வருகிற புது இலக்கண விளக்கங்கள் மற்றும் பண்பாட்டு அடிப்படையிலான விளக்கங்கள் அப்பாடத்திலேயே அவர்களுக்குக் கொடுக்கப்பட்டிருக்கும். உதாரணமாக http://learn.tamilnlp.com/unit_02/section_B/lesson01.html என்ற பக்கத்தில் தமிழகப் பண்பாட்டின் விருந்து பற்றிய விளக்கத்தை ஒளிக்காட்சி வாயிலாகக் கொடுக்கப்பட்டிருக்கிறது. விருந்தினராக ஒரு வீட்டுக்குச் செல்லும் போது அவர்களை நிறைய சாப்பிடச் சொல்லும் பழக்கம் தமிழர்களுக்கு உண்டு. அதே சமயத்தில் விருந்தினர்களும் தங்கள் வீட்டில் மனம்போன போக்கில் நிறைய சாப்பிடுவதையும் அவர்கள் விரும்புவதில்லை. இப்படியான சூழலில் தமிழ் மொழி வாக்கியங்கள் மற்றும் சொற்கள் ஆகியனவற்றை எப்படியெல்லாம் பயன்படுத்த வேண்டும் என்ற ஒரு நியதி இருக்கிறது. விருந்தினர் "போதும்! போதும்! வேண்டாம்! வேண்டாம்!" என அடிக்கடி சொல்ல வேண்டும். வீட்டினரோ! "இன்னும் கொஞ்சம்! இன்னும் கொஞ்சம்!" என அடிக்கடி சொல்ல வேண்டும். இப்படியான விருந்து மொழிப்பயன்பாட்டோடு பல்வகையிலும் தொடர்பு கொண்டிருப்பதைக் காணலாம். அமெரிக்கா போன்ற நாடுகளில் பிறந்த தமிழ்க் குழந்தைகளுக்கு இவ்வகையான பழக்கம் இருக்க வாய்ப்பில்லை! அவர்கள் வீட்டிலேயே அமரிக்கப் பண்பாட்டோடு வளர்ந்தவர்கள்! உணவும் உறவும் அவர்களின் விருப்பத்திற்கே விட

வேண்டும் எனும் கொள்கையில் வளர்ந்தவர்கள்! அவர்கள் தமிழகக் குடும்பத்தோடு தமிழகத்தில் கலக்கும் தருணத்தில் விருந்து போன்ற தமிழ்ப்பண்பாட்டை எதிர்கொள்ள வேண்டியிருக்கும்! இந்த வகையில் மொழியும் பண்பாடும் தமிழ்ச் சூழலும் பின்னிப் பிணைந்தவாறு கொண்ட பாடத்திட்டங்களை அமைத்துத் தருவது தமிழ்க்கற்பித்தலின் நோக்கமாக இருக்க வேண்டியிருக்கிறது. மேற்கூறிய பாடம் இந்தவகையில் அமைக்கப்பட்டதே! இங்கனமே இவ்விணையப்பக்கத்தில் கொடுக்கப்பட்ட ஒவ்வொரு பாடமும் தமிழ்ப்பண்பாடு, தமிழ்மொழிப் பயன்பாடு, மொழிச் சூழல் என்பனவற்றை மையமாகக் கொண்டு கொடுக்கப்பட்டிருக்கும்!

தமிழர்கள் புலம்பெயர்ந்து வேறு நாடுகளில் குடியுரிமைப் பெற்று வாழ்ந்து வருவதன் வரலாற்றை நோக்கும் போது புலம்பெயர்ந்த வம்சாவளி தமிழ் மாணவர்களுக்கான தமிழ்க்கல்விக்காகத் தமிழ் பயிற்றுவோர் அதிகக் கவனம் செலுத்த வேண்டியதன் முக்கியத்துவத்தைக் காணலாம். சிங்கப்பூர், மலேசியா, இலங்கை ஆகிய நாடுகளைத் தவிர மற்ற நாடுகளில் வம்சாவளித் தமிழ்க்கல்விக்கு அதிக முக்கியத்துவம் கொடுக்கப்படாத நிலையைத்தான் காணமுடிகிறது. மொரீஷியஸ், தென்னாப்பிரிக்கா போன்ற நாடுகள் மற்றும் சிங்கப்பூர், மலேசியா ஆகிய நாடுகளில் கிட்டத்தட்ட தமிழ்நாடு போன்று மொழிச்சூழல் ஈடுபாட்டுக்கான (language immersion) வசதிகள் இருக்கின்றன. ஆனால் அமெரிக்கா, ஐரோப்பா, கனடா போன்ற இடங்களில் இத்தகைய ஈடுபாட்டுக்கான வசதிகள் இல்லை! இத்தகைய ஈடுபாடு இல்லாத பட்சத்தில் தமிழகத்தில் கற்றுக்கொடுக்கப் பயன்படும் தமிழ்ப் பாடத்திட்டங்கள் எந்தவகையிலும் இவ்வகை நாடுகளில் பயன்படுத்த இயலாது. மாறாகத் தமிழ் மாணவர்களுக்குத் தமிழறிவு மட்டுமல்லாமல் தமிழ் உள்ளுணர்வு கொடுக்கும் வகையிலான தமிழ்ப்பாடத்திட்டத்தை வகுக்க வேண்டியதும் மிக அவசியம். முதல் நிலை தொடங்கி பத்தாம் நிலை, பதினொன்றாம் நிலையென தமிழ் வகுப்பைப் பிரித்துத் தமிழ்ப் பாடத்திட்டங்களை வகுக்கும் நிலை மாறி வம்சாவளித் தமிழ் மாணாக்கர்களின் இல்லத்தினின்று பெற்றத் தமிழறிவை மனதில் கொண்டு வெவ்வேறு வகையான பாடத்திட்டத்தை அமைப்பதில் தமிழ்க்கல்வியாளர்கள் கவனம் கொள்ள வேண்டியிருக்கிறது.

வம்சாவளி மாணாக்கர்களுக்கான தமிழறிவு புகட்டலில் இருக்கும் பல வருட அனுபவ அடிப்படையில் அவர்களை வயதை வைத்து வெவ்வேறு நிலைகளில் பிரித்துத் தமிழ்க்கல்வி கொடுக்க இயலாது என்பதை உணர முடிகிறது. இல்லத்தினின்று அவர்கள் பெற்ற தமிழ்க் கல்வியினடிப்படையில் அவர்களைப் பார்க்கும் போது ஒவ்வொரு மாணவரும் ஒவ்வொரு விதமான சூழலில் தமிழறிவு பெற்றிருப்பார்கள் என்பது தெரியவரும். இந்த வகையில் இவர்கள் அனைவரையும் ஒருங்கிணைய நோக்கி அவர்களது மொழியறிவை ஒரே தன்மையானது என கணிக்க இயலாது.

பெரும்பாலான வம்சாவளி மாணவர்களின் தமிழறிவின் தேவையைக் கணிக்கும் போது அவர்களுக்குத் தமிழ் உள்ளுணர்வு கொண்ட பேச்சுத் திறனும் எழுதும் திறனும் புகட்ட வேண்டியது மிக அவசியம் என்பதை அறியலாம். இவ்வகையில் இப்பகுதி மொழிச் செயல்முறைத் திட்டத்தில் எங்கனம் பாடத்திட்டத்தை அமைக்கலாம் என்பதை இதற்காக ஏற்படுத்தப்பட்ட தமிழ் இணையப் பக்கத்தின் அடிப்படையில் விளக்குகிறது. தமிழ் மொழி, மொழிச் சூழல்கள் மற்றும் தமிழ்ப் பண்பாடு ஆகியன ஒருங்கிணைந்த பாடத்திட்டத்தைத் தமிழ் வம்சாவளி மாணவர்களுக்கு அமைக்க வேண்டியதன் முக்கியத்துவத்தை இப்பகுதி விளக்குகிறது. tamilverb.com என்னும் இணையத்தளத்தில் ஆங்கில வினைகளின் அடிப்படையில் தமிழ் வினைகளைக் கொடுத்து அவை பயன்படுத்தப்படும் சூழல்களை விளக்கி அவற்றை ஒலிக்கோப்புகள் வழி கேட்கும் வசதியும் கொடுக்கப்பட்டிருக்கிறது. வம்சாவளி மாணவர்களில் பெரும்பாலோனோர் தங்களின் ஆங்கில மொழியின் அடிப்படையிலேயே தமிழ் மொழியைக் கற்பார்கள் என்னும் நோக்கில் அவர்களுக்கு ஆங்கில வினைகள் மூலம் தமிழ் வினைகளையும் அவற்றின் சூழல்களையும் அறிந்துகொள்வதற்கான வாய்ப்பு இவ்விணையப்பக்கம்.

2. கணினி வழி தமிழ் இலக்கிய ஆய்வு

தமிழ் இலக்கியத் தரவுகள் சங்க காலந்தொட்டு இக்காலம் வரையில் மின்வழியில் இருக்கும் இக்காலக்கட்டத்தில் அத்தரவுகளை முறையான தரவுத் தளத்தில் சேமித்து மிகவும் நுணுக்கமான வகையில் தேடும் சாதனங்களைப் பயன்படுத்தி அலச முற்படுவது என்பது இன்றியமையாத ஒன்றாகும். இம்முயற்சியில் பென்சில்வேனியாப் பல்கலைக்கழகத்தில் கீழ்க்காணும் இரு இணையப் பக்கங்கள் உருவாக்கப்பட்டிருக்கின்றன.[6]

6. இத்தளங்களில் கொடுக்கப்பட்டுள்ள தரவுகள் இணையத்தில் இலவசமாகக் கொடுக்கப்பட்டுள்ள மதுரைத் திட்டம், தேவாரம்.org, https://www.tamilvu.org/ ஆகிய தளங்களிலிருந்து பெறப்பட்டன. இத்தளங்களில் பயன்படுத்தப்படும் தேடுபொறி பி.எச்.பி நிரலி வழியாகவும் vue.js தொழிற்நுட்பம் வழியாகவும் கட்டுரையாளரால் உருவாக்கப்பட்டது. தரவுகள் MySQL தரவுத்தளத்தில் tamilnlp.com என்னும் தளத்தில் சேமிக்கப்பட்டுள்ளது.

sangam.tamilnlp.com





http://sangam.tamilnlp.com/mp/json/

இவ்விரு தளங்களின் முக்கிய நோக்கம் தமிழ் இலக்கியங்களிலிருந்து பாடல் வரிகளையும், சொற்களையும் சங்கம் முதல் இக்காலம் வரையிலான மொழி வரலாற்றின் அடிப்படையில் தேடும் வழியாகும். இத்தேடுபொறிகள் வழியாகப் பல சொற்கள் மற்றும் தொடர்களை அவற்றின் பயன்பாடு மற்றும் பொருள் அடிப்படையில் ஆய்வுசெய்ய வழிவகை செய்யலாம். எடுத்துக்காட்டாக 'முகில்' என்னும் சொல்லைத் தேடும்போது இச்சொல் பக்திகாலத்தில் அதிகமாகப் பயன்பாட்டில் இருப்பதையும் சங்ககாலத்தில் ஓரிரு இடங்களிலுமே வருவதைக் காணலாம். இச்செய்தியினின்று புலப்படும் ஆய்வு உண்மை என்னவெனில் பக்திக்கும் இச்சொல்லுக்கும் நெருங்கிய தொடர்பு இருப்பதை உணரலாம். சங்க இலக்கியங்களில் குறிப்பாக சூளாமணி, பரிபாடல் மற்றும் மணிமேகலை போன்ற பக்தி சார்ந்த இலக்கியங்களிலேயே

காணப்படுகிறது. பதிற்றுப்பத்து, முல்லைப்பாட்டு ஆகிய இலக்கியங்களில் ஒரு சில இடங்களிலேயே இச்சொல் வருகிறது. இது போன்ற ஆய்வு உண்மைகளை இத்தேடுபொறிகள் மூலம் எளிதாக அறியமுடியும். மேலும் பார்க்கல் ஆகும், சொல்லல் ஆகும் போன்ற அமைப்புகளையும் பார்க்கல் ஆமே, சொல்லல் ஆமே என்னும் அமைப்புகளையும் இத்தரவினின்று பார்க்கும்போது 'பார்க்கலாமே', 'சொல்லலாமே' என மருவி இடைக்காலத்தில் பயன்பாட்டில் வந்துள்ளது. இதையே மேலைநாட்டு மொழியியல் அறிஞர்கள் 'பார்க்க லாம்', 'சொல்ல லாம்' எனப் பிரித்து இலக்கணம் எழுதிவிட்டனர். 'பார்க்கல்', 'சொல்லல்' போன்ற அமைப்பு இக்காலத்தமிழில் மறைந்ததன் காரணம் இதுவே. இத்தகைய மிகவும் நுணுக்கமான தமிழ் வரலாற்றின் உண்மைகளை இத்தகைய தரவுகள் வழியே அறிய ஏதுவாகும். (இது குறித்த விரிவான விளக்கங்களுக்குக் காண்க அரங்கநாதன் 2023, 2020).

இவ்வகை ஆய்வு மேன்மேலும் வளம்பெற தமிழ்த் தரவுகள் அனைத்தையும் ஒரு தரவுத்தளத்தின் கீழ் கொண்டுவந்து செறிவான தேடுபொறிகளைக் கொண்டு பல கோணங்களில் தரவை அலச அனைத்து ஏற்பாடுகளையும் செய்ய வேண்டும். அதோடன்றி இத்தரவுத் தளங்கள் வழி செய்த ஆய்வை கட்டுரைகள் வழியாக மற்ற ஆய்வாளர்களுக்கு அறியப்படுத்த வேண்டும். இத்தகைய முயற்சிகள் தமிழ் இணையக் கல்விக்கழகம் (https://www.tamilvu.org/), செம்மொழித் தமிழாய்வு மத்திய நிறுவனம் (https://cict.in/cictinneww/) ஆகியனவற்றிலும் மேற்கொள்ளப்பட்டிருக்கிறது என்பது உண்மை.

3. தமிழ் பேசும் இயந்திரத்தை நம்மால் உருவாக்க முடியுமா? நமக்கிருக்கும் சவால்கள்.

தொழிற்நுட்பமும் செயற்கை நுண்ணறிவுத் திறனும் பலவித்திலும் வளர்ந்துவிட்ட இக்காலக்கட்டத்தில் தமிழ் தொழிற்நுட்பம் பற்றிய ஆய்வு பற்றி நாம் ஆழமாகச் சிந்திக்க வேண்டிய தருணத்தில் இருக்கிறோம். மனித இயந்திரத்தை உருவாக்கும் முயற்சியில் நாம் ஈடுபடும் தருணம் இது. தமிழ் பேசும் மனித இயந்திரம் மற்றும் தமிழில் நம்மோடு உரையாடும் இயந்திரம் என்பதுவே நமது குறிக்கோள். இம்முயற்சியில் இப்பகுதி இது குறித்தான ஆய்வில் ஈடுபடுகிறது. இப்பகுதியின் வழி ஒரு தமிழ் பேசும் இயந்திரத்தை இங்கு அறிமுகப்படுத்தவும் முயற்சிக்கிறோம். இப்பகுதி இயந்திரமும் தமிழும் என்னும் ஆய்வின் கட்டுரையாளரின் பல ஆண்டு முயற்சியில் உருவாக்கப்பட்டது. இது குறித்த ஆய்வின் முயற்சியை http://robot.tamilnlp.com என்ற முகவரியின் வழி நீங்கள் அறியலாம்.



இங்கு கொடுக்கப்பட்டுள்ள இணைப்புகளின் வழி இந்த இயந்திரம் பல செயல்பாடுகளை அறியலாம். அறிமுகம் குறித்த உரையாடல், ஆத்திசூடி பாடல்வரிகள், புறநானூறு வரிகள், தமிழ் மாதங்கள், போன்றவற்றை இந்த இயந்திரம் நம்மோடு இணைந்து செயல்படுகிறது. இதற்கான முக்கியக் கருவிகள் பேச்சுணரி மற்றும் உரையிலிருந்து பேச்சை மாற்றும் செயற்பாடுகள். இப்பகுதியில் தமிழ் தொழிற்நுட்பத்துக்கான ஆய்வுகள் மற்றும் செயற்திட்டங்கள் பற்றி விளக்க முற்படுவோம். தமிழ்க் கணினி ஆய்வை அடுத்தக் கட்டத்துக்கு நாம் எடுத்துச் செல்ல வேண்டும் என்பதுவே நமது முயற்சி. இவ்வாய்வு குறித்த விளக்கங்களை http://uttamam.org/papers/21_32.pdf என்னும் கட்டுரையில் விரிவாக அறியலாம்.

3.1. செயற்கை நுண்ணறிவுத் தொழிற்நுட்பம்:

செயற்கைத் தமிழ்த் தொழிற்நுட்பம் என்பது பற்றி நாம் சிந்திக்க வேண்டிய தருணம் இது. இயந்திரம் நம்முடைய அறிவுத்திறனைப் பெறும் வழிவகைகள் பற்றி ஆழமாகச் சிந்திக்க வேண்டும். நாம் எப்படி இயந்திரத்துக்கு நம் அறிவுத்திறனைத் தமிழ் வழி கொடுக்கப் போகிறோம் என்பதுவே நமக்கு இருக்கும் சவால். நம்மால் இது இயலுமா? நாம் நம்முடைய அறிவுத் திறன் பற்றி நன்கு அறிவோமா? இக்கேள்விகளுக்கு நம்மால் சரியாக விடை கூறமுடியாதுதான். இருப்பினும் ஒரு இயந்திரத்தை உருவாக்குவோமே. இவ்வியந்திரம் எப்படியெல்லாம் நம்மோடு உரையாடுகிறது என்று பார்ப்போம். இது ஒரு முன்முயற்சி. இம்முயற்சி வெற்றியடைந்தால்

நாமும் இயந்திரமும் ஒன்றாக வாழும் ஒரு சூழல் வர வாய்ப்பிருக்கிறது எனலாம்.

3.2. நமது முயற்சி:

இந்த இயந்திரத்தைப் பற்றிய விளக்கங்களை இங்கு விளக்குவோம். இது ராஸ்பெரி பை என்னும் ஒரு பட்டைக் கணினி (single board computer) வழி உருவாக்கப்பட்ட ஒன்று. இவ்வியந்திரத்தில் லினக்ஸ் என்னும் நிரலியின் வழிப் பல முயற்சிகளைச் செய்யும் வழி வடிவமைத்திருக்கப்பட்டிருக்கிறது. முக்கியமாக பைத்தான், பிஹெச்பி, மற்றும் மைசீக்குசவல் ஆகியவற்றின் திறனை முழுதுமாகப் பயன்படுத்தும் ஆய்வுத்திறன் என்று கூறலாம். இதோடு பேச்சைக் கேட்டு அதைக் கூகுள் பேச்சு உணரி மூலம் தமிழ் உரையாக ஆக்கும் முறையும் உரையிலிருந்து பேச்சுக்கு மாற்றும் முறையும் பைத்தான் நிரலி வழியாக செயல்படுத்தப்படுகிறது.

3.3. தமிழ் பேசும் நமது தமிழ்க் கருவி:

இக்கருவி நம்மோடு உரையாடுகிறது. இதற்குக் காரணம் கூகுல் வழிவகுத்த மொழியுணர் சாதனம் மற்றும் எழுத்திலிருந்து ஒலி, ஒலியிலிருந்து எழுத்து என்னும் பலவகை ஆய்வு நுட்பங்களை இங்கு உள்ளடக்கியிருக்கிறோம். இத்தகைய ஆய்வுமுயற்சிளோடு விக்கிப்பீடியாவிலிருந்தும் இலக்கியத் தரவுகளிலிருந்தும் செய்திகளைக் கொணரும் வழிமுறையையும் இக்கருவி ஒன்றுபட பயன்படுத்துகிறது. முக்கியமாகத் தமிழ்வழி இம்முயற்சி என்பதுவே இக்குறிக்கோள்.

3.4. இக்கருவியுடன் எப்படி நாம் உரையாடப் போகிறோம்?

இக்கருவி உங்களுடைய பேச்சைக் கவனமாகக் கேட்கப் போகிறது. நீங்கள் சொல்லும் வாக்கியங்களை நன்கு புரிந்துகொள்ளப் போகிறது. நீங்கள் சொன்னதைத் திருப்பிச் சொல்லப் போகிறது. எப்படி இது நடக்கிறது? இதுவே கூகிலின் இயந்திர ஆய்வின் வெற்றி எனலாம். பேச்சைப் புரிந்துகொள்ளும் அவர்களது நிரலி. அத்தகைய நிரலிகளை நமது நிரலியில் நாம் சரியாகப் பயன்படுத்திக்கொள்ள விரும்புகிறோம்.

3.5. விக்கித் தரவோடு தொடர்பு:

விக்கிப்பீடியா தமிழ்த் தரவோடு இணைக்கப்பட்டுப் பின்வரும் கேள்விகள் போன்ற பல கேள்விகளுக்கு இவ்வியந்திரம் நமக்கு விவரங்கள் தரும்.

நாம் கேட்பது: மயிலாடுதுறை பற்றி சொல்லுங்க:

இந்த இயந்திரம் சொல்வது: மயிலாடுதுறை (Mayiladuthurai) (முன்பு மாயவரம் என்று அழைக்கப்பட்டது) இந்தியாவில், தமிழ்நாடு மாநிலத்தில் மயிலாடுதுறை மாவட்டத்தில் உள்ள நிர்வாகத் தலைமையிடமும், சிறப்பு நிலை நகராட்சியும் ஆகும். மயில்கள் ஆடும் துறை என்பதால் மயிலாடுதுறை என அழைக்கப்படுகிறது.

3.6. கூகுள் மொழிபெயர்ப்பு நிரலியோடு தொடர்புபடுத்துவோம்

மின்வழி கூகுள் மொழிபெயர்ப்பு நிரலியோடு இணைக்கப்பட்டு இவ்வியந்திரம் பின்வருமாறு நம்மோடு உரையாடுகிறது.

ஹிந்தில சொல்லுங்க

हिंदी में कहें
hindee mein kahen

நீங்களும் உங்க தம்பியும் நாளைக்கு என்னோட வீட்டுக்கு சாப்பிட வரீங்களா

क्या तुम और तुम्हारा भाई कल रात के खाने के लिए मेरे घर आओगे?
kya tum aur tumhaara bhaee kal raat ke khaane ke lie mere ghar aaoge?

தமிழ் நல்லா பேச வருமா

क्या आप अच्छी तमिल बोल सकते हैं?
kya aap achchhee tamil bol sakate hain?

நீங்க எல்லாரும் இந்தியாவுல எந்தெந்த இடத்துக்கு போய் இருக்கீங்கன்னு ஒன்னு விடாம சொல்றீங்களா

क्या आप मुझे बता सकते हैं कि आप सभी भारत में कहां गये हैं?
kya aap mujhe bata sakate hain ki aap sabhee bhaarat mein kahaan gaye hain?

கனடால சொல்லுங்க

ಕೆನಡಿಯನ್ ಎಂದು ಹೇಳಿ
Kenaḍiyan endu hēḷi

நீங்க நாளைக்கு என்னோட வீட்டுக்கு வரீங்களா

ನಾಳೆ ನನ್ನ ಮನೆಗೆ ಬರುತ್ತೀಯಾ?
Nāḷe nanna manege baruttīyā?

நீங்க காலையில சாப்பாடு சாப்பிட்டீங்களா

ನೀವು ಬೆಳಿಗ್ಗೆ ತಿಂದಿದ್ದೀರಾ?

Nīvu beḷigge tindiddīrā?

கூகுள் மொழிபெயர்ப்பு வசதிப் பக்கத்தில் கொடுக்கப்பட்ட அனைத்து உலக மொழிகளையும் இந்திய மொழிகளையும் தமிழ் வழி மொழிபெயர்த்து அறியும் விதத்தில் இவ்வியந்திரம் வடிவமைக்கப்பட்டிருக்கிறது.

3.7. லினக்ஸ் சாதனத்தின் மணி, நாட்கள், மாதம் போன்ற வசதிகளைப் பயன்படுத்துவோம்.

லினக்ஸ் சாதனத்தில் உள்ள நேரம், காலம் ஆகிய வசதிகளைப் பயன்படுத்தி பின்வருமாறு இக்கருவி நம்மோடு உரையாடுகிறது.

நாம் கேட்பது: இப்ப மணி என்ன

இவ்வியந்திரம் சொல்வது: இப்பொழுது அமெரிக்கக் கிழக்கு நேரம் மாலை எட்டு மணி ஒன்பது நிமிடம் ஐம்பத்து ஒன்று வினாடி

நாம் கேட்பது: இன்னைக்கு என்ன தேதி

இவ்வியந்திரம் சொல்வது: இன்றைக்கு இரண்டாயிரத்து இருபத்து மூன்றாம் வருடம் பத்தாம் மாதம் பதின் ஒன்றாம் நாள்

எண்களைத் தமிழில் மாற்றி அவற்றைப் பேச்சில் மாற்றுவதற்குக் கட்டுரையாளரால் பைத்தானில் உருவாக்கப்பட்ட தமிழ் எண்களை மாற்றும் http://robot.tamilnlp.com/py/convert_tamil_number.py என்ற நிரலி இவ்வியந்திரத்தில் பயன்படுத்தப்படுகிறது. இவ்வாய்வு குறித்த மேலதிக விவரங்களுக்கு காண்க Renganathan 2023a.

4. தமிழ்க் கணினி ஆய்வின் எதிர்காலம்

இக்கட்டுரையில் விளக்கப்பட்டுள்ளபடி முக்கியமாக மூன்று துறைகளில் கணினியைப் பயன்படுத்துவது குறித்து நம் கவனத்தைச் செலுத்த வேண்டும். உலகலாவியத் தமிழர்களின் குழந்தைகளை மனதில் கொண்டு அவர்களுக்குத் தமிழ் கற்பிக்க http://learn.tamilnlp.com, https://www.tamilvu.org/ta/

கல்வி–விவரங்கள்–மழலைக்கல்வி போன்ற முயற்சிகள் தொடர வேண்டும். குறிப்பாக இம்முயற்சிகளைத் தொடர்ந்து பயன்பாட்டாளர்களின் பயன்பாடு வழி மதிப்பீடு செய்து இவற்றின் தரத்தைத் தொடர்ந்து வளப்படுத்த வேண்டும். இரண்டாவதாகத் தமிழ் இலக்கிய ஆய்வுக்கான sangam.tamilnlp.com, http://sangam.tamilnlp.com/mp/json/, http://tamilconcordance.in/ போன்ற இன்னும் பல தரவுத்தளங்களின் வளத்தைத் தொடர்ந்து செறிவுபடுத்த வேண்டும். குறிப்பாக தேடுபொறி வழி செய்யப்படும் ஆய்வுகளைக் கட்டுரையாகவும் நூலாகவும் வெளியிட்டுத் தமிழ் இலக்கிய ஆய்வாளர்கள் பலரை இவ்வாய்வில் ஈடுபடுத்த வேண்டும். மூன்றாவதாக robot.tamilnlp.com போன்ற தளங்களில் முயற்சி செய்யப்பட்டுள்ள கணினி நுண்ணறிவு ஆய்வுகள் வளப்படுத்த வேண்டும். https://bard.google.com/, https://chat.openai.com/auth/login போன்ற தளங்களின் வழித் தமிழுக்குக் கிடைக்கும் வளங்களின் அடிப்படையில் நமது தமிழ்க் கணினி ஆய்வுகளைத் தொடர்ந்து வழிநடத்த வேண்டும்.

துணை நூல்கள்

- Kawamura, Hiroaki. 2007. Participant Observation for Language Learners: A Performance-Based Approach to Language Learning during Study Abroad. Japanese Language and Literature. Vol. 41. No. 2. Oct. 2007.
- Miller, Cleve, 2008. Performance-based learning for teaching one-to-one classes (http://peo.cambridge.org/index.php?option=com_content&view=section&layout=blog&id=2&Itemid=8_)
- அரங்கநாதன், வாசு. 2023. தமிழ் மொழியின் வரலாற்றுப் பயணம்: சங்கம் முதல் இக்காலம் வரை காலச்சுவடு பதிப்பகம், நாகர்கோவில். (History of development of the Tamil language: From Sangam to modern period - in Tamil)
- Renganathan, Vasu 2023a. "Large Language Models (LLM) and the Role of the linguists in the World of AI" In The proceedings of the conference on செயற்கை நுண்ணறிவுத் தொழில்நுட்பத்தைத் தமிழ்மொழியுடன் செயல்படுத்துவதற்கான பன்னாட்டுக் கருத்தரங்கம், குமரகுரு கல்லூரி, கோயம்புத்தூர், October 13-14, 2023. (http://uttamam.org/papers/tic2023.pdf).
- Renganathan, Vasu 2020. 'Sangam to Modern Tamil Genre: The process of grammaticalization and evolution of Modern Tamil Noun and Verb forms.' IJDL:Vol. 49, No. 1 January 2020
- Renganathan, Vasu (2016). Computational Approaches to Tamil Linguistics, Cre-A. Chennai, India.
- Renganathan, Vasu (2016). Computational Approaches to Tamil Linguistics: Scopes and Prospects. In the proceedings of the 15th Tamil Internet Conference, Gandhigram Rural University, Dindigul, Tamil Nadu. (http://www.uttamam.org/papers/16_02.pdf).
- Renganathan, Vasu (2014). Computational Phonology and the Development of Text-to-Speech

Application for Tamil. In the Proceedings of the International conference on Tamil Internet, 2014, Pondicherry, India. (http://www.uttamam.org/papers/14_35.pdf).

- Renganathan, Vasu 2011. The Tamil Language in Context: A Comprehensive Approach to Tamil Learning. Department of South Asia Studies, University of Pennsylvania.
- Renganathan, Vasu 2009. Enhancing the processes of Learning Tamil with Synchronized Media, Paper presented at the 8th International Tamil Internet Conference, Cologne, Germany. (http://www.infitt.org/ti2009/papers/vasu_renganathan_final.pdf)
- Renganathan, Vasu (2001). Development of Morphological Tagger for Tamil, In the Proceedings of the International Conference on Tamil Internet 2001, Kuala Lumpur, Malaysia. (http://www.uttamam.org/papers/01_34.pdf).
- Renganathan, Vasu 1998. Formalizing the Knowledge of Heritage Language Learners: A Technology-Based Approach. Journal of the South Asia Language Pedagogy and Technology (http://salpat.uchicago.edu/index.php/salpat).

Concept Index: From Literary Study to Cultural Study

Dr. R. Jeyaraman

ABSTRACT

Concept index is a systematic collection of the all notions of a linguistic society. External objects including humans, relations between them and incidents and also thoughts, values, feelings, contexts of human being should be recorded in verbal form systematically in the concept index. Methods of language usage to be included in the index. This index includes both form of meaning and method of expression of meaning. It includes forming of the meaning and history of the meaning. Historically contemporary meaning of the words and modern meaning of the words both shall be analysed. Cultural meaning should be deconstructed and included in this index. Traditional meaning should be analysed through the knowledge of natural science and social science. Written language and spoken language are two different types. In spoken language sound, feelings of the speaker, context of the speech may have additional or negative meaning. These are supra segmental meanings or intonation meanings. Meaning is always beyond words. Extra meaning is between language and user of the language. Without the conscious knowledge of the speaker and writer meaning can be expressed regarding their gender, power, society, attitude and status. This meaning should be verbalized. Language has more meaning than words.

R. Jeyaraman
Thiruvalluvar University, Vellore.
jrjeyaramantvu@gmail.com

After parsing meaningful language categories can be available. Language categories can be classified and explained in grammar and lexicography. Noun meaning index is prepared based on word index for Sangam Literature and Thirukkural in Tamil. Dictionaries for all Sangam literatures with grammatical explanation are published by University of Kerala. The dictionaries were prepared based on noun and verb categories, did not considered the other categories like case, clitics, tense marker and PNG markers. Nonverbal categories are also more important in making meaning process. These categories are more associated with human mind and feelings. In literature and speech context meaning of the language is limited and pragmatic meaning and cultural meaning are unlimited. Pragmatic meaning and cultural meaning should be included in the concept index. Dictionaries, Meaning indexes and Encyclopaedia cannot cover the micro level important meanings. Meaning of the body language is also unique to include in the concept index.

A couplet in Thirukkural is taken for example to analyse for preparing concept index.

யாகாவா ராயினும் நாகாக்க காவாக்கால்
சோகாப்பர் சொல்லிழுக்குப் பட்டு (குறள். 127)

யா–கா–வ்–ஆ–ஆர் ஆ–ய்–இன்–உம் நா–கா–க்–க
கா–வ்–ஆ–க்–கால்
சோ–கா–ப்ப்–அர் சொல்ல்–இழுக்கு–ப் படு–உ

Parsing of the grammatical categories in the kural are given above. Two sentences are in the kural, couplet. Translation of the couplet is: “Whatever else you may control, control your tongue, lest you should repent your indiscreet words” (V. R. Ramahandra Dikkshitar 2000: 27). Sentence 1: Anyone may not control anything, but control speech behaviour. Sentence 2: If any one cannot control speech behaviour, he/she have to get blame of the people and to save the embracement situation. Link morph, increment, release vowels are important to describe structure of the language. Structure of the language is also concept.

Nouns – நா Tongue, சோ, சொல் word, இழுக்கு, யா interrogative pronoun

Verbs – கா control, protect, prevent (4), ஆ to become, படு auxiliary verb

This kural has cause and effect meanings. Cause is a previous event. Effect is a continuous event. Time is related in the cause and effect. Both meaning can be included in the concept index. Kaa is an intransitive verb. All nouns followed by accusative case marker and intransitive verb kaa should be collected and classified. Disease control, fort control, birth control, price control and other type of control to be collected and classified. These meanings can be included in the concept index. Prevent, protect, save like these verbs overlapped with control. Overlapping meanings can be analysed and included in the concept index. Controlling tongue has two meanings: 1. food control, 2. speech control. Kural says speech control in this context. All grammatical meaning are important for concept index. All types of noun followed by various case markers should be collected and classified. On the basis of cases the meanings of nouns to be explained in the concept index. bow, spear, vessel are instrumental meaning in the lexical category. The third case in Tamil has source, through, instrument and agent meanings. Finger, cloth, eye, book, brick can be instrument in sentences. These meanings are derived from grammatical categories. Accusative case, sociative case, ablative case, locative case and other such cases can determine the meaning of the nouns. Tense marker can determine the meaning of the action or verb and the agent of the action. A character may use future tense marker than other markers. Another character may use optative form of verbs than other verbs. Some characters may use polite form of words in their speech. Few characters may use negative verbs. In ethical literature more negative verbs are used by poets for valid reasons. These differences can help to identify the behaviour and concept of the characters. Tamil is an agglutinative language. Every bound morpheme has unique meaning.

Lexical meaning, cultural meaning and grammatical meanings are important to concept index. Two poems are taken to explain for preparing concept index.

சிற்றில் நற்றூண் பற்றி நின்மகன்
யாண்டு உளனோ என வினவுதி; என்மகன்
யாண்டு உளன் ஆயினும் அறியேன்; ஓரும்
புலி சேர்ந்து போகிய கல்அளை போல
தோன்றுவன் மாதோ, போர்க் களத்தானே.
(புறநானூறு. 86)

You grasp a fine pillar in my small house
You ask me "where is your son?"
I do not know where he is.
Like a mountain cave that a tiger
In-habited and abandoned,
is this womb which gave birth to him.
He will appear on the battle field. – (vaidheki)

அண்ணாந்து ஏந்திய வனமுலை தளரினும்,
பொன்நேர் மேனி, மணியின் தாழ்ந்த
நல்நெடுங் கூந்தல் நரையொடு முடிப்பினும்
நீத்தல் ஓம்புமதி பூக்கேழ் ஊர!
இன்கடும் கள்ளின் இழைஅணி நெடுந்தேர்க்
கொற்றச் சோழர் கொங்கர்ப் பணீஇயர்
வெண்கோட்டு யானைப் பேளர் கிழவோன்
பழையன் வேல் வாய்த்து அன்ன நின்
பிழையா நன்மொழி தேறிய இவட்கே. (நற்றிணை. 10)

Even when (her) beautiful breasts, looking
upwards (and) eminent, slacken
on her who has come to believe in your unfailing
good words,
like in the success of the spear of Palaiyan,
lord of Poor with white-tusked elephants,
when the Konkars were humbled by the
victorious Coolars
with jewl-adorned long chariots (and)
sweet strong toddy,
(and) if (she) ties, because of whiteness,
(her) good long tresses
which hang down without jewels on (her)
gold-like body
-beware of leaving (her), o man from village
rich in flowers.
– (Eva Wilden 2008: 79)

Literary meaning is emerging from Linguistic meaning. Word order in the sentence, selection words of the sentence may have different meaning from normal meaning. Author's gender, social, cultural, regional and historical back-rounds have different meanings. Author's intentions can have some other meanings. Figure of speech may have internal meanings. All meanings can be included in the concept index. Each literary genre have some specific method of meaning expressions. The above mentioned two poems from Sangam literature belong to agam and puram genres. Knowing literary tradition is essential to make concept index. A women/girl asked the neighbour women mother where about of the later's son. The mother answered about her son warrior. This is the content of the poem (puram. 86). Feelings of the mother can be included in the concept index. Proud, confident, comparison and doubt are important to this index. Mother- son relation and citizen – king relations, war activity of the king, role of the female citizen are also important.

In the agam poem (naRRinai. 10) loyal girl friend of the heroine told some advices to the hero about

heroine's well future. She is comparing between present young love relation and old age life relation. Conduct of the hero should not be changed. He should not lose his oath. In this poem through the aesthetic point of view describes the body parts of the young heroine and in the future of old heroine. Breast, hair and skin of the lady are described in the poems. Through the description these are the aesthetic concepts of the old sangam period. These concepts are not changed still now. Possibility of the departure of the hero from heroine is alive in her girlfriend mind. Status of the women in ancient society is depicted in this poem. Feminist and post-modernism point of views this may have extended meanings. Poet's intention is to praise Palaiyan, the king. Through the simile poets include the king's war skill in the poem.

For more interpretations in both modern and old literary texts need literary and art theories. Textual theories, narratology, discourse, rhetoric, structuralism, de-construction, semiotic are some of the theories for helpful to make concept index. In computer analysis these type of meanings cannot drawn through word index. Key word index is little helpful. Words are the way of all the meanings. Word indexes can be prepared through drawing, photo then videos, movies. In future events directly converted into word forms. Now face recogniser through CCTV cameras, lie detector, speech to written form converter, written form to speech converter are available in the technology field.

Concept index can have more meanings than dictionary, meaning index and encyclopaedia. Language computing can be helpful to this concept index.

BIBLIGRAPHY

- Ewa Wilden (Editor and Translator), 2008, NaRRinai 1-200, A Critical Edition and an Annotated Translation of Narrinai Volume I, Puducherry: Ecole Franaise D'Extreme- Orient and Chennai: TamilMan Pathippakam.
- Ramachandra Dikshitar V. R. (Translator), 2000, Tirukkural, Chennai: The Adyar Library and Research Centre
- https://songamtranslationsbyvaidehi.com

MACHINE TRANSLATION AND MULTILINGUAL TECHNOLOGIES

Domain Adaptation of Bidirectional Neural Machine Translation system involving Tamil to Telugu

Yash Bhaskar, Nagaraj V, Vandan M, Dipti Misra Sharma, Parameswari Krishnamurthy

ABSTRACT

Machine Translation (MT) has evolved significantly, and the development of domain-specific systems has become pivotal in improving translation accuracy. Constructing effective domain-specific machine translation systems requires a comprehensive understanding of domain terms and their precise translation. This article explores the significance of domain term identification, and building domain adapted Tamil-Telugu machine translation systems. In the context of Tamil and Telugu, experiments are conducted using a neural machine translation (NMT) system, IndicTrans2, across five domains. The experiments show that fine-tuning NMT with domain terms significantly enhances the translation quality.

Yash Bhaskar, Nagaraj V, Vandan M, Dipti Misra Sharma, Parameswari Krishnamurthy

International Institute of Information Technology, Hyderabad

yash.bhaskar@research.iiit.ac.in,
nagaraju.vuppala@research.iiit.ac.in,
vandan.mu@research.iiit.ac.in,
dipti@iiit.ac.in,
param.krishna@iiit.ac.in

1. INTRODUCTION:

Machine Translation (MT) has witnessed significant advancements in recent years, with the development of domain-specific systems playing a crucial role in enhancing translation accuracy and relevance. Building domain-specific machine translation systems requires a thorough understanding of domain terms and their accurate translation. This article delves into the importance of domain term identification for constructing effective machine translation systems involving Tamil and Telugu, prominent Dravidian languages.

The term 'domain' refers to a specific situation or a state in which a group of people share a common knowledge on a particular subject matter. The domain terminologies in each domain are restricted within its domain boundary differentiating one domain from the other. A domain term refers to a word or phrase specific to a particular industry, field, or domain. Domain term identification is a critical step in the development of domain-specific machine translation systems. Identifying and translating these terms accurately is essential for ensuring that the machine translation system produces linguistically and contextually relevant output in specialized domains. In essence, domain term identification involves systematically interpreting word meanings in both the source and target languages within the context of a particular domain. It helps in the identification of terms specific to the domain and greatly contributes to NLP applications and in the language translation process.

2. CHALLENGES IN DOMAIN TERM TRANSLATION

Generic translation models may struggle to capture the intricacies of domain-specific content, leading to inaccurate and contextually inappropriate translations. There are confusions among researchers with domain terms and named entities. Domain terms and named entities are both essential linguistic elements, but they differ in their scope and significance within the context of natural language processing and machine translation. Domain terms encompass words or phrases specific to a particular industry, field, or domain, serving as key identifiers within specialized vocabularies. These terms

contribute to the precision and fluency of translations in domain-specific contexts. On the other hand, named entities refer to specific objects, individuals, locations, or organizations that hold unique identities. While named entities can be domain-specific, they extend beyond specialized terminology to include proper nouns and entities with distinct characteristics. Both domain terms and named entities play crucial roles in the accurate representation of information, and their identification is integral to developing effective language models, machine translation systems, and other natural language processing applications.

As domain terms are specialized terminologies related to the field, translation of the same into the target language is likely to face challenges, especially in cases where the terminologies exhibit Polysemous and homonymous meanings. While the former deals with words having multiple senses that are related, the latter deals with meaning senses that is unrelated. While evaluating domain terms in the given translations, we came across semantically unrelated words in the target language in each domain in terms of polysemy and homonymy. This is mainly due to the fact that Polysemous words in one language do not find their direct mapping in other languages in general.

Consider the example given in Table-1 for homonymous word forms in English translated into Tamil and Telugu. The term intelligence is a homonymous word which has three distinct meanings in different domains in Tamil and Telugu.

Domain	Explanation	English Sentence	Tamil Sentence	Telugu Sentence
Cognitive Science	Mental capacity for learning, reasoning, problem-solving.	Intelligence is crucial for success in academics.	கல்வியில் வெற்றிபெற **அறிவுத்திறம்** முக்கியமானது.	విద్యావిషయక విజయానికి **తెలివితేటలు** చాలా ముఖ్యమైనవి.
Artificial Intelligence	Simulated intelligence in machines and computer systems.	Artificial intelligence is revolutionizing various industries.	செயற்கை **நுண்ணறிவு**, பல்வேறு தொழில்களில் புரட்சியை ஏற்படுத்தி வருகிறது.	కృత్రిమ **మేధస్సు** వివిధ పరిశ్రమలను విప్లవాత్మకంగా మారుస్తోంది.
Military	Gathering and analyzing information for defense purposes.	The Intelligence Bureau is India's domestic intelligence arm.	**உளவுத் துறை** என்பது இந்தியாவின் உள்நாட்டு உளவுப் பிரிவாகும்.	**ఇంటెలిజెన్స్ బ్యూరో** అనేది భారతదేశపు దేశీయ గూఢచార విభాగం.

Table 1: Homonymous domain terms

Besides homonymous word forms, Polysemous word forms as well pose challenges in the domain term translation process. Consider the examples given in Table-2 for the word depression in English and the equivalent word form in Tamil and Telugu in different domains.

Domain	Explanation	English Sentence	Tamil Sentence	Telugu Sentence
Clinical Science	State Feelings of persistent sadness and hopelessness.	Kumar has been struggling with **depression** for several months.	குமார் பல மாதங்களாக **மன அழுத்தத்துடன்** போராடி வருகிறார்.	కుమార్ కొన్ని నెలలుగా డిప్రెషన్ తో బాధపడుతున్నాడు
Psychological Disorder	A serious mental health condition.	The doctor diagnose her with clinical **depression**.	மருத்துவர் அவளுக்கு மருத்துவ **மனத்தளர்ச்சி** இருப்பதைக் கண்டறிந்தார்.	వైద్యుడు ఆమెకు క్లినికల్ **మానసిక ఒత్తిడితో** బాధపడుతున్నట్లు నిర్ధారించారు.
Economics	Prolonged economic decline with high unemployment.	The country is experiencing a severe economic **depression**.	நாடு கடுமையான பொருளாதார **மந்தநிலையை** சந்தித்து வருகிறது.	దేశం తీవ్ర ఆర్థిక **మాంద్యం** ఎదుర్కొంటోంది.
Physics and Meteorology	Area of low atmospheric pressure.	A low-pressure **depression** is moving towards the coast.	குறைந்த காற்றழுத்த **தாழ்வு பகுதி** கரையை நோக்கி நகர்கிறது.	**అల్పపీడనం** తీరం వైపు కదులుతోంది.
Chemical Science	Lowering of freezing point due to solute addition.	The addition of a solute causes a **depression** in freezing point.	ஒரு கரைசலைச் சேர்ப்பது உறைபனி புள்ளியில் **அழுத்தக்குறைவை** ஏற்படுத்தும்.	ఒక ద్రావణాన్ని కలపడం వలన గడ్డకట్టే ప్రదేశంలో **ఒత్తిడి** ఏర్పడుతుంది.

Table 2: Polysemous domain terms

As some of the language properties such as polysemy and homonymy are inherent to any natural language, it is necessary to disambiguate words with multiple meanings for the right interpretation of its sense in relation to the context. Since the domain term identification process facilitates systematic interpretation of word meaning, it is employed in disambiguating Polysemous and homonymous domain words. Therefore domain term identification is considered as an indispensable process through which word interpretation can be effectively achieved in their respective domains which in turn helps in accurate translation of the domain terms into the target language.

3. DOMAIN TERM IDENTIFICATION IN TAMIL-TELUGU

Domain term identification ensures precision in translating content within a specific domain. Various efforts are put into building domain-specific MT systems for different languages (Arcan, M., et al., 2014; M. Amin Farajian et al., 2017; Hu, J., et al. 2019;

Elise Michon et al., 2020). It is found to be necessary to build a domain aware machine translation system in the context of Tamil too. Terms and expressions that hold domain-specific meanings play a pivotal role in conveying nuanced information accurately.

We conducted experiments to investigate how a neural machine translation system handles domain-specific translation for low-resource language pairs, specifically Tamil to Telugu. We constructed small yet significant domain corpora for Tamil-Telugu language pairs across five domains. The details of the developed domain corpora, including sentences, tokens, and domain terms for each domain, are presented in Table-3

Domain	#Sentences	#Tokens (Tamil / Telugu)	#Domain Terms (Unique Count / Total Count)
Agriculture	5014	49846 / 50821	5097 / 12113
Education	5182	54490 / 56558	1752 / 3758
Government	2769	35010 / 34410	3848 / 9814
Health	2507	23977 / 23740	2724 / 6034
Science	2684	29861 / 29973	5007 / 11537
Total	18156	193184 / 195502	16138 / 43256

Table 3: Domain Data Developed for Tamil-Telugu

Additionally, we created a total of 500 sentences for development and testing data across these domains. Our experiments were carried out using the state-of-the-art system IndicTrans2 (Gala, et al. 2023). We report the BLEU ((Papineni, K. et al. 2007) and ChrF scores (Popović, M., 2015) for the Tamil-Telugu translation direction on the Flores test set ((Costa-jussà, et al. 2022) as 17.7 and 49, respectively.

4. DOMAIN ADAPTATION OF TAMIL-TELUGU NEURAL MACHINE TRANSLATION SYSTEM

We utilized the aforementioned training data to customize the indictrans2-indic-indic-dist-320M model. Employing the LoRA (Low-Rank Adaptation) technique for fine-tuning, we efficiently adapted pre-trained models to specific tasks or domains. LoRA accomplishes this by freezing the pre-trained model weights and introducing trainable rank decomposition matrices into each layer of the Transformer architecture, thereby significantly reducing the number of trainable parameters for downstream tasks. In our experiments, we conducted a single LoRA adaptation for all these domains, resulting in a substantial reduction of the number of trainable parameters and GPU memory requirement. The LoRA parameters used for our fine-tuning are detailed in Table-2.

Parameter	Value
lora_target_modules	q_proj,k_proj,v_proj
lora_dropout	0.05
lora_r	8
lora_alpha	32

Table 4: LoRA parameters for our fine-tuning

5. RESULTS

We assessed the performance of our model using four different methods: BLEU, ChrF, Comet (Rei, R., 2020) and a novel metric that examines domain term coverage in the translated text.

Given that we have already annotated domain terms in domain corpora, we verify whether the marked terms on the target side in the reference data are present in the model's translation. This coverage is denoted as a percentage in Table-5.

Model Type	Bleu	CHRF	COMET	Domain Term coverage (%)
Baseline (indictrans2-indic- indic-dist-320M)	5.397	37.9849	0.7574	21.8442
Fine-tuned (5 Epoch)	8.337	43.8809	0.7497	24.17072
Fine-tuned (10 Epoch)	8.142	44.9808	0.7513	26.59594

Table-5: Overall Performance

These results suggest that extended fine-tuning positively impacts the model's overall performance, reflected in improved scores across the evaluated metrics. Notably, the novel metric addressing domain term coverage offers valuable insights into the model's proficiency in incorporating domain-specific terminology into its translations.

6. CONCLUSION AND FUTURE WORK

Any Machine Translation system that is trained on huge data, learns from the vast variety of data. But it is of no guarantee that it performs on all domains. In this paper by human and automatic evaluations it is evident that domain adaptation is needed and can improve the results. The main advantage of domain adaptation is the MT system that can be fine tuned to train on any domain provided there are domain terms. While MT systems trained on extensive data offer versatility, domain adaptation is crucial for improved performance across diverse domains. The inclusion of domain terms in MT systems facilitates learning specific domains, enhancing translation quality. Future work may focus on building domain adapted machine translation systems using domain terms involving Tamil and other languages.

Acknowledgements:

This work is supported by the project “Discourse Integrated Dravidian Language to Dravidian language Machine Translation (DiscoMT)” under National Translation Mission (NLTM): BHASINI

REFERENCES

- Arcan, M., Giuliano, C., Turchi, M. and Buitelaar, P., 2014.. Identification of bilingual terms from monolingual documents for statistical machine translation. In Proceedings of the 4th International Workshop on Computational Terminology (Computerm) (pp. 22-31).
- Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J. and Sun, A., 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.
- Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In Proceedings of the 28th International Conference on Computational Linguistics, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Gala, J., Chitale, P.A., AK, R., Doddapaneni, S., Gumma, V., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V. and Kumar, P., 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. arXiv preprint arXiv:2305.16307.
- Hu, J., Xia, M., Neubig, G. and Carbonell, J., 2019. Domain adaptation of neural machine translation by lexicon induction. arXiv preprint arXiv:1906.00376.
- Hu, J., et al. 2021 "Lora: Low-rank adaptation of large language models." arXiv preprint arXiv:2106.09685.
- M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In Proceedings of the Second Conference on Machine Translation, pages 127–137, Copenhagen, Denmark, September. Association for Computational Linguistics
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. 2002.. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).
- Popović, M., 2015, September. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the tenth workshop on statistical machine translation (pp. 392-395).
- Rei, R., Stewart, C., Farinha, A.C. and Lavie, A., 2020. COMET: A neural framework for MT evaluation. arXiv preprint arXiv:2009.09025.

A Novel Input Method for Tamil

Baskaran Sankaran

ABSTRACT

The present input methods for Tamil, while providing robust support for using the language in Computers, mobile and other devices, have several shortcomings. We explore the current input methods and discuss their shortcomings in detail. We seek to design a new Input Method for Tamil that overcomes these issues and provides a consistent and seamless experience across different devices with varying form factors.

Unlike existing keyboards, we consider the Phonemic nature of the Tamil script by using the vowels and pure consonants as the base units. We performed an extensive statistical analysis of Tamil characters from large Tamil corpora to understand their frequencies. This analysis together with the language heuristics were then used to design an optimal keyboard layout across dominant and weaker finger positions to enable faster input and to reduce the finger movements in touch typing.

Finally, we compare our proposed Tamil input method with existing approaches and show its advantages over the others in an objective manner.

Baskaran Sankaran

Maadhyamik Technologies
baskaran@maadhyamik.com

1. INTRODUCTION

Presently there are 3 predominant input methods that are in wider use.

- Anjal-style transliteration
- Tamil99
- Google Keyboard (GBoard)

The inadequacy of existing Input methods for Tamil has been studied several years ago (Sendhil Kumar Cheran et al. 2004). More recently, Elango Cheran (2022) expanded on this and published a detailed blog post in expounding his idea of exploiting the phonemic nature of the Tamil alphabets for the new Input method.

We take a similar approach and design a new Input method. However, unlike Cheran's proposal, we design our keyboard layout based on the frequency analysis of Tamil characters from a large corpora. This strategy enables us to decide the key position for Tamil characters so that frequent characters can be placed on the dominant finger positions. Secondly, unlike some existing layouts, ours is fully compatible with the QWERTY layout for the placement of punctuations and symbols, thereby enabling easier, seamless transition between typing in Tamil and English.

Several studies (Fagarasanu and Kumar 2003; Naomi G. Swanson et al. 2018; Thomsen et al. 2008; Van Tulder et al. 2007; Wright and Atkinson 2019) has investigated the occupational hazards in work settings and has determined some link between the typing in workplace and Repetitive Strain Injury (RSI) or even Carpal Tunnel Syndrome (CTS).

Some studies (Erich Grunewald 2022; Naomi G. Swanson et al. 2018; Shieh and Lin 1999) has also focussed on different keyboard layout designs by considering several typing related metrics such as hand alternation, load distribution between both hands, same hand roll combos, weaker (pinky/ little) finger. We consider these metrics throughout our design effort in trying to develop an optimized keyboard layout with better ergonomics, while dealing with varying constraints imposed by these metrics. To our knowledge, this is the first work that looks into these factors in designing a input method for Tamil.

Overall, this enables faster typing and at the same time reduces typing stress/ fatigue on fingers. We also evaluate our new phonemic layout with two existing input methods on some empirical metrics.

2. TAMIL INPUT METHODS: SHORTCOMINGS

There are several disadvantages to the existing Tamil input methods.

Unnatural Design: In Tamil Consonant Vowels such as ‘க’ and ‘தை’ are generated by the combination of pure consonants ‘க்’ and ‘த்’ with the vowels ‘அ’ and ‘ஐ’ respectively. Thus the Vowels and the Consonants, which form the basic units of the sounds (phonemes) in Tamil, should be the basis of designing a good Input method. Several Input methods including the Tamil99 follow the unnatural design of Vowels and Consonant vowels (CV) such as ‘க’, ‘ங’ and ‘ச’, as the basic units in the Keyboard. To be fair, this design came to be used, because these CV characters. One of the unintended consequence of this design is that this can produce illegal character sequences in Tamil such as vowels followed by vowel modifiers (such as ‘் ‘pulli) eg. அ் or with dangling consonant vowel modifiers.

Transliteration Dependency: Anjal and other transliteration keyboards presume users to be familiar with the English alphabets and require them to transliterate the Tamil sounds to one or more English character sequence. The user will actually be inputting the Tamil words in transliterated English letters, which are then mapped back to Tamil by the keyboard engine. This method is hugely popular primarily due to the high English literacy among the Tamil speaking population around the world. However, we believe this is doing more harm because the new speakers no longer have to learn the script but only the sounds in the language. Another disadvantage with this approach is that there are multiple ways to represent a Tamil character in English, because of the variations in the sounds in the two languages.

Non-adherence to QWERTY layout: Most of the Tamil keyboards do not adhere to the widely-used QWERTY layout in terms of the key positions reserved for punctuations and other symbols in the keyboard. These Input methods assign Tamil characters in these positions. Consequently, the bilingual users using QWERTY will find it difficult to switch back and forth between English and Tamil typing and they will be forced to learn and follow the different key positions for typing punctuations and symbols while using Tamil.

GBoard Design Incongruity: The Google Keyboard or GBoard for Tamil is the soft key layout launched for touch devices. It lays out the Vowels (அ, ஆ... ஃ) and Consonant Vowels (‘அ’ வரிசை CVs such as க, ங, ச..., ஷ, ஹ) in a 9x4 matrix. The vowel characters panel on the left changes every time a consonant vowel is pressed to show its other CV variations. The layout uses a simplistic sequential positional of characters in the alphabet, without any concern for either optimizing finger movements or the character frequency based layout design. Combined with the incongruity of ever changing vowel panel, GBoard’s design choice is probably the least efficient Tamil key layout in use. Further this layout is limited to the touch interfaces and may not be readily adapted for keyboard based input.

3. DESIGNING A NEW INPUT METHOD

3.1 Design Principles

We wanted to design a new Input Method for Tamil that address the shortcomings in the existing one and also make it easier to learn the new method with a short learning curve. Based on our research, we decided on the following design goals for the new Input method.

1. A design that adheres to and exploits the Phonemic nature of Tamil, taking the phonemes as the basic unit
2. Frequency analysis of base phonemes and consonant vowel combination in order to achieve an optimal design that speeds up touch typing in computers and equivalently reduces finger movement in touch devices
3. Intuitive arrangement of keys to make the learning easier that is consistent across different platforms and devices
4. Prevent any illegal character sequences in the output text
5. Maximize compatibility with the QWERTY keyboard to make the transition between English and Tamil typing seamless and easier.
6. Eliminate the forced requirement for the user to know other script/ language and instead facilitate typing in the Tamil script

Further, we seek to carefully consider typing related factors such as hand alternation, load distribution between both hands, same hand roll combos in our design. As far as we know, this is the first work that looks into some of these aspects in designing a input method for Tamil.

It should be noted that, the same design principles could be used for designing better Input methods for other Abugida languages as well.

3.2 Tamil Dataset

We started the design by identifying Tamil corpus data for doing usage frequency analysis of the characters in the language. We identified large enough corpora (approx. 537M words) from mainly two different sources as below:

- Kaggle - Tamil Language Corpus for NLP [1]
  - Tamil Articles Corpus
  - Tamil New Corpus
  - Tamil Language Corpus

- Github - Opensource Tamil Corpus [2]
  - Tamil Wikipedia
  - The Hindu Tamil corpus

3.3 Frequency Analysis

Our goal is to understand the usage frequency of basic phonemes as well as for the full set of Tamil alphabets: vowels, consonants and the consonant vowel (CV) combinations. Once we understand the usage frequency of the phonemes and the full set of alphabets, we can exploit this information to design the keyboard layout. It should be noted that we have omitted the Sanskritized characters (வடமொழி எழுத்துகள்) ஜ், ஶ, ஷ etc. from this analysis.

Figure 1 shows the frequency analysis of all Tamil characters across 3 categories: vowels, consonants and consonant-vowels (CV). Among the top-10 most frequent characters, we have 5 consonant vowels and 4 'அ' ending CVs and

1. Consonants: ம், ர், ல், க் and ன் - 349.13M
2. 'அ' ending CVs: க, த, ப and வ - 306.19M
3. 'உ' ending CV: து - 64.63M

Thus by using the base phonemes (vowels and pure consonants) for our keyboard layout, would result in a saving of nearly 43M keystrokes for this dataset. Now consider two more heatmaps i) by characters ending with vowel sounds (column-wise sum of the above heatmap) in Figure 2 and ii) by the characters for each consonant-vowel series (row-wise sum) in Figure 3.

Notice that the pure consonants (right-most cell) tend to be more frequent than any consonant vowel series. Here again by using the basic phonemes as the keys instead of the 'அ' ending consonant vowels, these pure consonants can be typed with a single key press as opposed to two presses, saving about 37M keystrokes on this dataset. Also notice that the vowels and CVs ending in short form vowel sound are much more frequent than their long form counterparts.

Using the frequency statistics of the Vowels (first two in the first heatmap) and the Consonant vowels above, we can design optimal keyboard layout to minimize the movement of fingers and to use the dominant fingers for the high frequency phonemes. The next section discusses the design decisions and explains our Tamil Phonemic keyboard layout.

4. PHONEMIC KEYBOARD LAYOUT DESIGN

As we mentioned earlier in our design goals, we want the new keyboard layout to be easier for the users to learn and use across different devices with varying form factors. Given the constraints of available keys (in QWERTY layout) and total required keys to accommodate Tamil characters and symbols, we had to make certain design decisions in the character placement to the different key positions within the layout.

1. We want to use home row of the keyboard for the Vowels as well as some high-frequency Consonants in Tamil.
2. We want to use home row of the keyboard for the Vowels as well as some high-frequency Consonants in Tamil.
3. We also believe that the key positions corresponding to the dominant fingers (index and middle fingers) in two non-home rows should take precedence over the home row

1. https://www.kaggle.com/datasets/praveengovi/tamil-language-corpus-for-nlp
2. https://github.com/ajithalbus/TamilCorpus

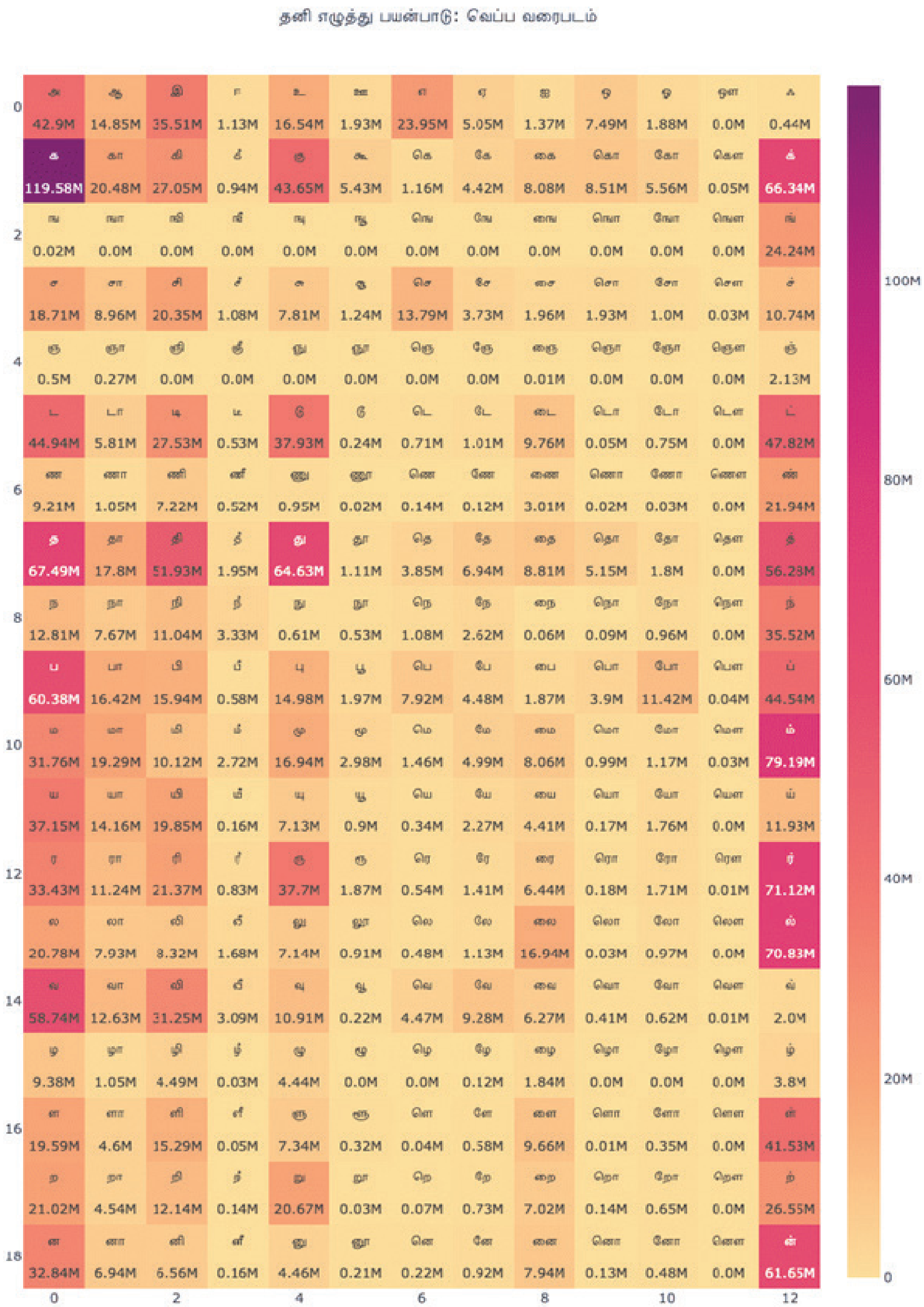


Fig-1: Tamil Characters - Frequency Statistics Heatmap

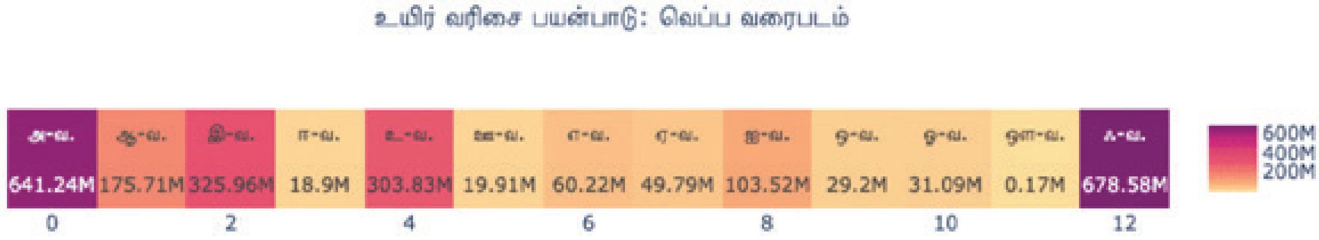


Fig-2: Tamil Vowels - Frequency Statistics Heatmap

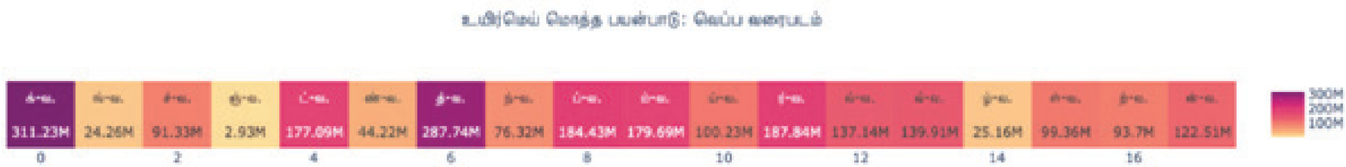


Fig-3: Tamil Consonants - Frequency Statistics Heatmap



Fig-4. Our proposed Keyboard Layout for *regular* (top) and *shift* (bottom) faces

keys with weaker fingers. We'll be using this later in optimizing the character assignment to the key positions. Given the high frequency of vowel characters (in standalone or CV forms), we have retained most of the vowels (except for ஐ and ஔ) in the left hand-side of the *home* row. The frequent vowel அ has been assigned to the dominant left-hand finger position (key position 'G' in QWERTY) and the other short form vowel letters are assigned in the right to left order in the same row. The corresponding long

form vowels have been assigned to the same keys in the Shift row. Both these factors together, allow for easier key position recognition for the users thereby reducing the cognitive load for the users in trying to find the keys.

Several other Tamil keyboard layouts have also assigned the vowels on the left side. However, unlike the other layouts, our assignment is based on the usage frequency and finger strength. For example, most layouts including the Tamil99, has assigned அ to the key position of 'a' in QWERTY. However, given its very high frequency, this adds significantly more load on the left-hand weaker (pinky) finger.

In the next step, we have allocated most of the consonants are going on the dominant right-hand side of the keyboard given their high frequency. Thus, for typing the most of the CV combinations, the user will be using both hands, thereby enabling hand alternation. This allows the CVs to be typed efficiently by a mix of both hands, without making the same hand/ finger to move to a different position for typing a single character.

Based on our observation #2 above, the frequent Consonants starting with 'க்' and 'த்' are assigned to the dominant finger positions in the home row and the rows above and below. We assign the rest of the consonants to the successively weaker key positions in the decreasing frequency order.

We then specifically considered the case of மெல்லினம் (nasalized consonants), which are typically be followed by the corresponding வல்லினம் (plosive/ stop consonant) in Tamil. Thus, it made sense for us to place these nasalized consonants on the left side of the keyboard (above and below the home row) so that the following வல்லினம் can be typed with the right hand. This encourages hand alternations for நட்பெழுத்து combinations (such as ங்க, ஞ்ச etc.), which occur frequently in the Tamil corpora.

We made an exception for 'ம்' and assign it to the dominant key position on the right side, due to its high frequency in both Consonant and CV forms.

Some of the frequent bigram CCV forms include consonant combinations of ர்ச், ம்ப், ன்ற் etc., which involve hand rolls (multiple keys typed with the same hand in a single movement) thereby reducing frequent hand alternation.

On the Shift key layout, we assigned the Tamil numerals right below the roman numerals to make the typing intuitive and easier. Additionally, the Sanskritized consonants and other Tamil symbols are assigned on this Shift layout.

Figure 4 above shows our proposed Tamil keyboard layouts for the regular and shift faces. Our design also achieves reasonable balance between the load/ fatigue on both hands, making it easier for the average users and particularly more so for the left-handed users.

We now present some analysis of our keyboard layout against two well-known input methods using few objective metrics.

4.1 Typing Effort: # of Keystrokes

Based on the keyboard layouts for the three input methods, viz Anjal transliteration, Phonemic and Tamil99, we analyzed their efficiency and ease of typing in two ways. We first calculated the number of absolute keystrokes required to type the Tamil words in the above corpora of 537M words used in this work. To keep the analysis simple, we ignored the punctuations and any non-Tamil words/ characters for this. We also ignored the shift key here because the shift key is pressed simultaneously with the key following it. Here are the absolute number of keystrokes required for typing the above Tamil corpora by the 3 input methods.

- Anjal: 4,470,795,879 (4.47 B)
- Tamil99 : 4,124,838,873 (4.12 B)
- Phonemic: 4,045,040,635 (4.04 B)

Our new Phonemic method requires the least number of keystrokes among the 3 methods we use for comparison; specifically it requires 80M fewer keystrokes than Tamil99. This is because, the pure consonants are usually frequent than their CV combination. In contrast, Tami99 layout requires an additional keystroke: ‘ ் ‘ (pulli) for inputting pure consonants. For Anjal, we used the standard transliteration mapping as suggested in the Sellinam app, thus requiring two keystrokes for each long vowel as well as for long CV combinations.



Fig-5. Typing effort on different key positions: Anjal (top), Tamil99 (middle) & Phonemic (bottom)

4.2 Typing effort on different key positions

We now analyze the heatmap on the keyboard layouts of the three input methods to see which keys are typed in more frequently and their relative position in the keyboard. We plotted the heatmap on the 3 keyboard layouts separately for this analysis. As above, we ignored the punctuation marks and non-Tamil words to keep this analysis simple. However, we considered the shift key in this analysis.

We can understand that, a layout will be easier for typing if the frequently typed characters are i) in the position of dominant fingers of either hands or ii) in the home row of the keyboard. The Anjal keyboard is clearly the least efficient option as the most of the frequently used keys are outside of the dominant finger positions of the keyboard.

As shown in Figure 5, unlike in Tamil99 layout the frequent characters are mostly placed in the dominant finger positions in our new phonemic keyboard, which makes the typing easier. The dominant left and right index fingers positions (in all 3 rows) alone account for 61.22% and 46.32% of the overall typing in Phonemic and Tamil99 keyboards respectively. This difference of 15% is significant and makes the Phonemic layout a better (in terms of ease of use) option than the Tamil99 keyboard.

We then look at the percentage of typing for the keys in the home row, which is the usual resting position for the hands when not typing. It thus has the advantage that the user will not have to move their hands from its resting position. If we only consider the home row typing, our new phonemic layout accounts for 58.33%, while the Tamil99 is slightly better with 61.83% of

overall typing. We believe this small difference of home row typing is far outweighed by the advantage gained in the Phonemic keyboard layout by the dominant index fingers across all the rows. In addition to the efficiency in typing, the new Phonemic keyboard layout offers other advantages over the Tamil99 keyboard as discussed earlier.

5 CONCLUSION

We have proposed a new phonemic-inspired keyboard layout for Tamil typing which has strong advantages over existing Tamil input methods. Our keyboard layout offers consistent design and interface on devices of varying form factors for wider and seamless adoption.

We have recently launched the தமிழ் இழை (Tamil Izhai) keyboard apps in Android[3] and iOS[4] platforms based on our new phonemic layout. We are currently working to support new slide-typing feature to make Tamil input even easier and faster on mobile devices. While, our apps might be relatively new, we believe that the users will gradually realize the benefits of the optimized keyboard layout and the consistent, seamless experience across different devices/platforms.

REFERENCES

- Carlos Ignacio Jr. Lugay, Yoshiki Kurata, Joseph Ramon Leofando, Janfil Mari Pamisal, and Jeryl Salas. 2022. An Analysis on the Effects of Different Types of Keyboards on Users' Productivity and Hand Muscle Strain. In 13th International Conference on Applied Human Factors and Ergonomics (AHFE) 2002.
- Elango Cheran. 2022. Redesigning an Input Method for an Abugida Script.
- Erich Grunewald. 2022. How Bad Is QWERTY, Really? A Review of the Literature, such as It Is.
- Fagarasanu, M., & Kumar, S. (2003). Carpal tunnel syndrome due to keyboarding and mouse tasks: a review. International Journal of Industrial Ergonomics, 31(2), 119-136
- Naomi G. Swanson, Traci L. Galinsky, Libby L. Cole, Christopher S. Pan, Steven L. Sauter. 1998. The impact of keyboard design on comfort and productivity in a text-entry task. Applied Ergonomics. Volume 28, Issue 1.
- Sendhil Kumar Cheran, Thuraiappah Vaseeharan and Elango Cheran. 2004. Optimization of Tamil Phonetic Keyboard. In Tamil Internet Conference. 2004.
- Shieh, K. K., & Lin, C. C. 1999. A quantitative model for designing keyboard layout. Perceptual and motor skills, 88(1), 113-125.
- Thomsen, J. F., Gerr, F., & Atroshi, I. 2008. Carpal tunnel syndrome and the use of computer mouse and keyboard: a systematic review. BMC musculoskeletal disorders, 9(1), 1-9.
- Van Tulder, M., Malmivaara, A., & Koes, B. 2007. Repetitive strain injury. The Lancet, 369 (9575), 1815-1822.
- Wright, A. R., & Atkinson, R. E. 2019. Carpal tunnel syndrome: An update for the primary care physician. Hawai'i journal of health & social welfare, 78(11 Suppl 2), 6.

1. https://t.ly/OAdk
2. https://t.ly/r5k7

Design and Development of a Neural Machine Translation System for Kannada – Tamil

Dr. B.Ashwath Rao

ABSTRACT

Kannada and Tamil are two of the Dravidian Languages spoken majorly in the South Indian states of Karnataka and Tamil Nadu. Both these two languages are scheduled and classical languages of India. Though both these two states are neighbors there is very little work done in the translation of text, and speech from one language to another. Machine translation refers to the use of computer algorithms and technology to automatically translate text or spoken words from one language to another. The goal of machine translation is to facilitate communication between people who speak different languages by providing quick and reasonably accurate translations. Rule-based Machine Translation, Statistical Machine Translation, Neural Machine Translation, and Transfer Learning in Neural Machine Translation are some of the techniques in Machine Translation. As part of a funded project titled "Discourse Integrated Dravidian Language – Dravidian Language Machine Translation Kannada-Tamil", we have developed a bi-directional Neural Machine Translation system. The system uses sequence a sequence mapping technique to translate from the source language to the target language. The tokenized word is encoded first and then transferred using a decoder. The system also uses an attention mechanism to focus on the region in the input sequence during learning. We have used Byte Pair Encoding (BPE) in encoding the source language text. We have trained the system on parallel corpus in a computing environment with a Graphical Processing Unit. We have also trained further on the publicly available parallel corpus. The total number of sentences in the training corpus is 24,98,652. The system is evaluated using BLEU, ChrF2, TER, and COMET evaluation metrics. We have obtained a BLEU score of 5.0, ChrF2 score of 36.2, TER score of 83.2, and COMET score of 0.7438 on Flores Test data in Kannada-Tamil translation. On the other hand, in Tamil-Kannada translation on Flores test data, we obtained a BLEU score of 4.4, ChrF2 score of 33.0, TER score of 83.7, and COMET score of 0.6655. This is the first neural machine translation system from Kannada to Tamil to the best of our knowledge

Dr. B. Aswath Rao,
Associate Porfessor, MIT, Manipal , Karnataka.
ashwath.rao@manipal.edu

INTRODUCTION

The Dravidian language family is one of the world's major language groups, primarily spoken in Southern India and certain regions of Sri Lanka. The Dravidian language family is known for its linguistic diversity, with around 80 different languages and dialects. The major Dravidian languages include Tamil, Telugu, Kannada, and Malayalam, spoken by millions of people across South India. Predominantly spoken in Southern India, Dravidian languages have a significant presence in the states of Tamil Nadu, Andhra Pradesh, Karnataka, and Kerala. Tamil, in particular, has a rich literary tradition and is one of the oldest living languages in the world. Dravidian languages are characterized by agglutination, where prefixes and suffixes are added to the root words to convey meaning. The languages exhibit a subject-object-verb sentence structure, in contrast to the subject-verb-object structure of Indo-Aryan languages. Tamil, one of the classical Dravidian languages, has a history dating back over two millennia and boasts a vast body of literature, including the Sangam literature. The Dravidian language family has contributed significantly to Indian culture, philosophy, and art. Each major Dravidian language typically has its own script. Tamil uses the Tamil script, Telugu uses the Telugu script, Kannada uses the Kannada script, and Malayalam uses the Malayalam script. Beyond India, the Dravidian influence extends to Sri Lanka, where Tamil is spoken as a major language. Dravidian languages have also influenced the cultural and linguistic landscape in Southeast Asia. Dravidian languages continue to play a vital role in contemporary India, both socially and politically. Efforts are ongoing to promote and preserve these languages, including initiatives for education and cultural enrichment.

Machine Translation (MT) can be defined as an automated system that analyzes text from a Source Language (SL), applies computations to that input, and produces equivalent text in a required Target Language (TL), ideally without any human intervention (Koehn, 2010). It is one of the most interesting and challenging problems in the field of Natural Language Processing (NLP). The two primary challenges in machine translation are adequacy and fluency. Adequacy focuses on developing a system that accurately represents the ideas expressed in the source language into the target

language. Fluency, on the other hand, emphasizes representing those ideas grammatically.

Kannada and Tamil are two of the Dravidian Languages spoken majorly in the South Indian states of Karnataka and Tamil Nadu. Both these two languages are scheduled and classical languages of India. Though both these two states are neighbours there is very little work done in the translation of text, and speech from one language to another. Machine translation refers to the use of computer algorithms and technology to automatically translate text or spoken words from one language to another. The goal of machine translation is to facilitate communication between people who speak different languages by providing quick and reasonably accurate translations. Rule-based Machine Translation, Statistical Machine Translation, Neural Machine Translation, and Transfer Learning in Neural Machine Translation are some of the techniques in Machine Translation.

In the rule-based approach, the text in the source language undergoes analysis using various tools, such as a morphological parser and analyzer, which transform it into an intermediate representation. A set of rules is then employed to generate the text in the target language based on this intermediate representation. A substantial number of rules are essential to encompass the complexities of natural language. These rules serve to transfer the grammatical structure from the source language to the target language. However, as the number of rules increases, the system becomes more intricate (Islam et al., 2010) and tends to slow down during translation. The formulation of a large number of rules is a tedious process, requiring years of effort and linguistic analysis.

Statistical Machine Translation (SMT) involves the automated conversion of sentences from one human language, the source (e.g., French), to another human language, the target (e.g., English). This process is conceptualized as a stochastic, or probabilistic, system. Various SMT variants exist, and they differ in how translation is modeled. These approaches include string-to-string mapping, trees-to-strings, and tree-to-tree models. Despite their differences, all these variants share the fundamental idea of automating translation. Models are trained using parallel corpora (pairs of source and target sentences) and monolingual corpora (examples of target sentences) (Osborne, 2011).

Neural Machine Translation (NMT) is an approach that employs neural networks for the task of machine translation. Unlike traditional methods that involve distinct steps and components, NMT utilizes a single neural network to perform translation. NMT relies on deep neural networks, specifically sequence-to-sequence models. The architecture consists of an encoder and a decoder network that work together to transform input sequences in the source language to output sequences in the target language. Instead of breaking down the translation process into separate components (such as alignment and decoding), NMT allows for end-to-end learning. The entire translation model is trained jointly, optimizing the network's parameters to minimize translation errors. Learning-based translation models in NMT utilize sequence-to-sequence models. These models are designed to handle variable-length input and output sequences. The encoder processes the source sequence, creating a fixed-size context vector that captures the input's semantic information. The decoder then generates the target sequence based on this context vector. NMT models are trained using parallel corpora, which consist of pairs of source and target sentences. The neural network learns to map input sequences to output sequences by adjusting its parameters during the training process. NMT represents words and phrases as continuous vectors in a high-dimensional space, capturing semantic relationships and contextual information. This enables the model to learn more nuanced and context-aware translations.

Transfer Learning in Neural Machine Translation (NMT) involves utilizing pre-trained models to enhance the performance of translation tasks. Transfer learning in NMT entails leveraging knowledge gained from pre-training on one language pair or domain and applying it to improve the performance of a related translation task. The primary goal is to make efficient use of pre-existing knowledge, reducing the need for extensive labeled data in the target domain and improving the convergence speed during fine-tuning. There are two Strategies: a) Multi-Lingual Pre-training: Models are trained on data from multiple languages, enabling the network to learn universal linguistic features that can be fine-tuned for specific language pairs. b)Single-Language Pre-training: Models are initially trained on a large corpus from a specific language pair, and this pre-trained model is fine-tuned for a related language pair.

As part of a funded project titled "Discourse Integrated Dravidian Language – Dravidian Language Machine Translation Kannada-Tamil", we have developed a bi-directional Neural Machine Translation system. The system uses sequence a sequence mapping technique to translate from the source language to the target language. The tokenized word is encoded first and then transferred using a decoder. The system also uses an attention mechanism to focus on the region in the input sequence during learning. We have used Byte Pair Encoding (BPE) in encoding the source language text.

LITERATURE REVIEW

There have been a lot of approaches for machine translation. Earliest approaches for machine translation

include rule-based. In this approach, hand-crafted rules are provided to system as a reference guide for the system in translation. This involves the need of language expert. However, to attain high accuracy from these systems, an exhaustive rule has to be set up. The next approach is Statistical Machine Translation (SMT). In this, statistical measures like probability is used in predicting the target word corresponding to source word [4]. Brown et al., have formulated a method for translation using parallel corpus. It involves two steps. In the first step, distribution of likelihood of target sentences are derived to form P(t) using a Language Model. in the next step, the probability of source sentence given target sentence P(s|t). The maximum value of these values is found.Neural machine translation is a corpus based approach in machine translation. It is ideal for sentence level sequence to sequence mapping. It uses attention based encoder-decoders. It is further extended with self-attention based transformers[4]. Suppose S is a source text and T is the translated text, then S is broken into s1, s2, …sn, where each si is a sentence. In the encoding stage, a series of fixed vector S1, S2, …SN is generated.

$$P(T|S) = P(T|\ S1,S2,\dots\ SN) \qquad (1)$$

In the decoding stage, each word is predicted based on the corresponding encoded vector Si and source sentence vector.

$$P(T|S) = P(ti\ |\ t0,t1,t2,,,,ti\text{-}1;\ s1,s2\dots sN) \qquad (2)$$

Bi-RNNs are a pair of RNNs, for processing in both directions. One RNN for processing in forward direction and another for processing in the backward direction. Transformer is a self-attention based model architecture with multihead mechanism [4]. This has gained popularity and uses encoder-decoder pair

From the visible research there is no neural based system developed for Kannada-Tamil or Tamil-Kannada translation. However there are about 5 web/mobile applications for translation.

METHODOLOGY

A Transformer model is trained to perform Neural Machine Translation (NMT). Samanantar dataset (Ramesh et al, 2021) is used to train IndicTrans model for 11 Indic languages. This also includes Kannada and Tamil. It is a multilingual NMT model, and it transliterates all text into Devanagari script for sharing lexical details among language pairs Kannada and Tamil for transfer learning. It also prevents word fragmentation in case of subword vocabulary and allows usage of a smaller subword vocabulary. This model uses 400M parameters.

In our NMT system, IndicTrans architecture used has six encoder-decoder layers, input embeddings size of 1536 with 16 attention heads and feedforward dimension of 4096 with a total of 434M parameters. Here Adam optimizer (P and Kingma, 2014) is used, label smoothing value is set to 0.1, and gradient clipping is 1.0 with learning rate: 0.0005. Warm-up steps are 4000 and cross entropy loss is considered.

The table 1 gives the data distribution sentence wise across train, validation, and test sets.

Table 1: Data split for training the NMT model

Data	Count of sentences (Kannada-Tamil)	Source
Train	26,04,203	Samanantar +Kanaja
Validation	1000	Samanantar
Test	1012	FLORES

The NMT system is trained on Kannada Tamil parallel corpora with 2604203 sentences from Samanantar parallel corpus (Ramesh et al, 2021) and Kanaja (Chilume, 2023) with a split of 24,98,652 sentences from Samanantar and 1,05,551 sentences from Kanaja. For validation, the Benchmark data of AI4Bharat is used which has 1000 sentences. And for testing, 1200 sentences from FLORES data (Facebook, 2022) are used.

RESULTS

The translation results seem promising and can be improved further by enhancing the dataset and the training phase. The table 2 and table 3 gives the evaluation scores on the test set for Kannada to Tamil and Tamil to Kannada NMT models.

Table 2:
Evaluation scores for Kannada to Tamil NMT system

Evaluation metric	Score
BLEU	6.4
ChrF2	39.6
TER	79.3
COMET	0.7879

Table 3:
Evaluation scores for Tamil to Kannada NMT System

Evaluation metric	Score
BLEU	6.0
ChrF2	36.6
TER	80.8
COMET	0.7322

The BLEU(Kishore,2002), ChrF2 (Maja, 2015) and COMET(Ricardo, 2020) scores are comparatively higher for Kannada Tamil language pairs whereas TER score is higher for Tamil to Kannada translation which indicates Kannada Tamil translation quality is better than Tamil Kannada. BLEU (BiLingual Evaluation Understudy), is a string-based automatic metric where as COMET (Crosslingual Optimized Metric for Evaluation of Translation) uses token/sentence embeddings to obtain similarity between the translated output and a reference translation.

Though these metrics correlate, different values are obtained due to the different formulas of each metric. Lower BLEU score does not imply that translation is completely incorrect, as it is an automated metric which uses n grams precisions and focusses on string similarity it may be misleading. Whereas chrF (CHaRacter-level F-score) relies on n-gram F score. chrF calculates the similarity between a machine translation output and a reference translation using character n-grams and is hence better than BLEU.

TER (Translation Edit Rate) is an error-metric for machine translation. It is calculated by the number of edits required to change a translated output into one of the reference sentences. It is generally preferred over BLEU for assessing sentence post-editing effort. Lower scores indicate better translation as the number of post edits to the translation is minimum.

REFERENCES

- Philip Koehn. 2010. A book on Statistical Machine Translation. Cambridge University Press.
- Islam, Zahurul, Jörg Tiedemann, and Andreas Eisele. 2010. English to Bangla phrase-based machine translation. In Proceedings of the 14th Annual conference of the European Association for Machine Translation.
- Osborne Miles 2011. Statistical Machine Translation. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_783
- Asha Hegde, Hosahalli Lakshmaiah Shashirekha, Anand Kumar Madasamy, and Bharathi Raja Chakravarthi. A Study of Machine Translation Models for Kannada-Tulu. 2022. In Congress on Intelligent Systems, pp. 145-161. Singapore: Springer Nature Singapore.
- Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. Transactions of the Association for Computational Linguistics, 10, 145-162.
- Diederik P and Jimmy LeiBa Kingma. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
- Facebook research . 2022. Facebookresearch/Flores: Facebook Low Resource (FLORES) MT benchmark, GitHub. Available at: https://github.com/facebookresearch/flores (Accessed: 04 January 2024).
- Chilume– ಚಿಲುಮೆ. 2024. Available at: http://chilume.com/?cat=9 (Accessed: 04 January 2024).
- Kishore Papineni , Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 311-318.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Rajaram BS, Ramakrishnan AG, Kumar HS. 2013. An accessible translation system between simple Kannada and Tamil sentences. In Proceedings of 6th language and technology conference.

Machine Translation: A Comprehensive Survey

E. Sivakumar, R. Anitha

ABSTRACT

Machine translation, the automated process of translating text from one language to another, has witnessed significant advancements in recent years. This review paper aims at providing a comprehensive overview of the developments in machine translation techniques and its impact on language translation tasks. An extensive review of the literature is carried out on various approaches to machine translation, covering traditional rule-based methods, statistical machine translation, and the emergence of neural machine translation architectures, analyzing their strengths and limitations. The review discusses various challenges in machine translation for Indian languages due to their strong morphology and limited resources. The paper highlights the evaluation metrics, the impact of machine translation on various domains, including business, education, and cross-cultural communication. By synthesizing the existing literature, this review paper aims to provide valuable insights into the current state of machine translation, its challenges, and future directions

E. Sivakumar,
Research Scholar, Department of CSE, Sri Venkateswara College of Engineering, Sriperumbudur.

R. Anitha,
Professor, Department of CSE, Sri Venkateswara College of Engineering, Sriperumbudur.

1. INTRODUCTION

The field of machine translation (MT) has undergone remarkable transformations since its inception. The roots of machine translation can be traced back to the mid-20th century when researchers embarked on the ambitious task of developing automated systems to translate human languages. One of the pioneering efforts was the development of rule-based systems that relied on linguistic rules and dictionaries to perform translation.

The limitations of rule-based approaches became evident as they struggled to handle the nuances, idioms, and context-dependent nature of human languages. Despite these challenges, these early endeavors laid the groundwork for subsequent advancements in the field.

As computational capabilities improved, statistical approaches gained prominence in the 1990s. Statistical Machine Translation (SMT) marked a departure from rule-based systems by leveraging probabilistic models and large bilingual corpora for training. SMT demonstrated notable success in capturing the statistical patterns of language pairs, allowing for more context-aware translation.

The last decade has witnessed a paradigm shift with the advent of Neural Machine Translation (NMT). NMT models, based on artificial neural networks, have demonstrated unprecedented accuracy and fluency in translation tasks. The rise of deep learning techniques, particularly the use of recurrent and transformer architectures, has propelled NMT to the forefront of machine translation research.

The journey from rule-based systems to statistical models and, more recently, neural network-based approaches showcases the resilience and adaptability of the field in addressing the intricate challenges posed by natural language. This survey aims to explore the diverse methodologies that have shaped the landscape of machine translation.

The objectives of this survey are multifaceted, aiming to provide a comprehensive exploration of the landscape of machine translation. The key goals include (i) undertaking a thorough examination of existing literature to capture the evolution of machine translation methodologies and pivotal research contributions that

have shaped the field (ii) offering insights into the various methodologies employed in machine translation and their strengths and limitations (iii) identifying the challenges inherent in machine translation for Indian languages and (iv) highlighting the recent advancements in machine translation, with a specific focus on emerging trends and breakthroughs

2. METHODOLOGIES IN MACHINE TRANSLATION

This section discusses about the existing methodologies in Machine Translation.

2.1 Rule Based Machine Translation

Rule-Based Machine Translation (RBMT) represents one of the earliest approaches to automated language translation. In RBMT systems, linguistic rules and dictionaries are meticulously crafted to define the mapping between source and target languages. These rules encode syntactic, semantic, and morphological structures, allowing for a systematic transformation of input sentences into the desired output language.

M. f. Alawneh et al. (2013) proposed a combination of RBMT and Example Based Machine Translation (EBMT) techniques for English-Arabic pair and reported an average precision of 97.2%. Sinha et al. (1995) developed a rule based translation system for English-Indian languages. Their approach used a interlingua model that used an intermediate representation of source language prior translating to target form

ArviHurskainen and Jorg Tiedemann (2017) developed a rule based automated translation for English-Finnish pair by employing syntactic and semantic rules. The authors focused on the inflection forms of both languages. However, they mentioned that it is difficult to frame rules for all scenarios.

In summary, Rule-based Machine Translation, with its foundational principles and ongoing innovations, represents a crucial chapter in the evolution of machine translation

2.2 Statistical Machine Translation (SMT)

As RBMT systems evolved, research efforts focused on addressing the limitations of manual rule creation. Statistical Machine Translation (SMT) represents a paradigm shift in machine translation by leveraging statistical models to automatically learn translation patterns from parallel corpora, marked a departure from rule-based systems and introduced data-driven methods for handling the complexities of language translation.

Brown et al. (1990) laid the groundwork for statistical machine translation by introducing the IBM Models. These seminal models formed the basis for aligning words in parallel corpora and estimating translation probabilities.Koehn et al. (2003) introduced the phrase-based statistical machine translation. This paper outlined a framework where translations occur at the level of phrases rather than individual words, improving the handling of long-range dependencies. The authors demonstrated their model over six European language pairs

Philipp Koehn and Hieu Hoang. (2007) introduced Factored Statistical Model, an extension of Phrased Based model. They stated that factoring can be obtained by reorganizing the words at sentence level prior translation, yields better results. Och and Ney (2004) made significant contributions to decoding strategies in statistical machine translation. Their work introduced efficient search algorithms, such as beam search, for finding optimal translations within the vast search space.

Koehn and Knight (2002) developed a lexicon for German-English pair from monolingual corpora over specific context. They reported 39% accuracy of noun translation. Foster and Kuhn (2007) addressed the challenge of domain adaptation in statistical machine translation. Their work focused on adapting translation models to specific domains, improving the relevance of translations in specialized contexts.

Koehn and Monz (2006) explored and compared manual and automatic methods for evaluating the quality of machine translation systems between six European languages to arrive at baseline translation scores.

In summary, Statistical Machine Translation has witnessed significant developments over the years, with foundational models evolving to address diverse challenges.

2.3 Neural Machine Translation (NMT)

Neural Machine Translation (NMT) has emerged as a transformative paradigm in machine translation, leveraging artificial neural networks to model the complex relationships between source and target languages. The architecture of neural networks allows for the end-to-end learning of translation patterns, enabling NMT models to capture intricate linguistic dependencies.

Sutskever et al. (2014) proposed sequence-to-sequence (seq2seq) architectures for neural machine translation using Recurrent Neural Networks (RNNs) to generate translations in an end-to-end manner. The Seq2Seq model struggled with long input sequences.

Bahdanau et al. (2014)introduced the attention mechanism, a key innovation in NMT. The attention mechanism allows the model to focus on different parts of the source sentence during the translation process, improving the handling of long-range dependencies. The attention mechanism addressed the limitation of fixed-size context vectors. This innovation allowed the

model to focus on different parts of the input sequence during translation, significantly improving the handling of long-range dependencies.

Wu et al. (2016) presented Google's Neural Machine Translation (GNMT) system. This influential work showcased the scalability of NMT models and their ability to outperform traditional statistical machine translation systems.

The initial NMT model considered word level input for translation which suffered understanding the semantics. K. Chen et al. (2019) proposed a NMT model with Convolutional Neural Network (CNN) and attention mechanism that considers sentence level input to improve context-based translation.

3. RECENT ADVANCES IN NMT

The ability of deep learning models to automatically learn hierarchical representations of input data has proven crucial in capturing the complexities of language translation NMT has witnessed significant advancements in recent years, with the emergence of transfer learning techniques and the transformative impact of transformer-based models.

3.1 Transformer Models

Transformer-based Neural Machine Translation (NMT) systems have become the de facto architecture, outperforming traditional approaches in terms of both accuracy and efficiency. The introduction of transformer architectures has revolutionized machine translation, offering superior performance, scalability, and the ability to capture long-range dependencies.

Luong et al. (2015) extended the attention mechanism by proposing global and local attention models. This work refined the attention mechanism, enhancing the model's ability to align and translate source and target language sequences effectively.

Vaswani et al. (2017) introduced the transformer architecture, a pivotal development in NMT. The Transformer architecture has become a cornerstone in machine translation models. The transformer model leverages self-attention mechanisms to capture contextual information, leading to significant improvements in translation accuracy and training efficiency.

3.2 Transfer Learning

Transfer learning has gained prominence in MT, enabling models to leverage knowledge gained from one task or language pair to improve performance on another. This approach is particularly beneficial in low-resource scenarios.

Devlin et al. (2018) introduced a pre-training approach for deep bidirectional transformers. Bidirectional Encoder Representations from Transformers (BERT) to enhance language understanding. BERT employs transformer architecture with multiple layers, enabling bidirectional processing of input data. The model is pretrained on large amounts of unlabeled text using unsupervised learning. BERT introduces a novel training objective called the Masked Language Model, where random words in a sentence are masked, and the model is trained to predict those masked words based on the context provided by the surrounding words.

Devlin et al. (2019) introduced mBERT (Multilingual BERT), extending BERT, addressing the need for a model that can effectively handle multilingual tasks without being fine-tuned for each language separately.

3.3 Zero-Shot Machine Translation

Zero-shot translation refers to the capability of a machine translation system to translate between language pairs for which it has not been explicitly trained. The model learns to capture underlying patterns, smenatics, and representations that are useful for translation tasks, allowing it to generate reasonable translations for language pairs not explicitly encountered during training.

Johnson et al. (2017) extended the capabilities of NMT to zero-shot translation. The authors demonstrated the capability of a single NMT model to translate between multiple languages, language pairs for which no parallel training data was available, demonstrating its generalization capacity.

3.4 Multi-Modal Translation

Multimodal translation refers to the process of translating content that includes multiple modalities or types of information, such as text, images, and possibly other forms of data like audio.

Lu et al. (2019) extended deep learning models to handle multimodal translation, incorporating visual information alongside textual input. This work signifies the growing trend of integrating multiple modalities for more context-aware and comprehensive translation

4. MACHINE TRANSLATION FOR INDIAN LANGUAGES

Machine Translation (MT) in Indian languages has garnered significant attention due to the linguistic diversity across the subcontinent.

4.1. Review of Machine Translation research for Indian Languages

A. Gupta, B. Patel (2020) implemented a neural machine translation system for major Indian languages. They reported that they achieved state-of-the-art performance, demonstrating the effectiveness of neural

models. Their work highlighted the importance of domain-specific data for improved translation accuracy. They mentioned that the performance was limited by the availability of parallel corpora for certain low-resource languages

S. Kumar, M. Verma (2015) developed a rule-based MT system for Hindi-English translation and mentioned that the RBMT was efficient for handling morphological challenges in Hindi. They emphasized the role of linguistic rules in addressing language-specific nuances. However, the approach suffered limited scalability for languages with complex syntactic structures

R. Menon, S. Nair (2018)applied statistical models to translate South Indian languages. They demonstrated the adaptability of SMT to diverse language families. The authors stressed the need for comprehensive linguistic resources for accurate translation. However, they found that SMT is sensitive to data quality, especially for low-resource languages.

N. Das, A. Ghosh (2017) combined rule-based and statistical methods for English to Bengali translation and shown improved translation quality by leveraging both linguistic rules and data-driven models, showcasing the potential synergy of hybrid approaches in bridging translation gaps. However, the approach was complex and faced maintenance challenges

P. Joshi, R. Kapoor (2019)utilized deep learning for code-switching translation in multilingual Indian contexts. They have successfully handled language mixing, a common phenomenon in Indian language, identifying the need for models capable of understanding and preserving code-switching nuances. They mentioned that there are challenges in fine-tuning for specific domains.

K. Rao, S. Desai (2021) investigated machine translation challenges for the low-resource language Kannada and shed light on the importance of resource augmentation for underrepresented languages. They advocated for collaborative efforts in building linguistic resources for improved translation.

M. Singh, A. Jain (2016) explored cross-lingual transfer learning for improving Hindi translation and provided insights into effective strategies for cross-lingual transfer learning. They showcased the potential of leveraging knowledge from resource-rich languages. They mentioned that the dependency on the availability of parallel data for source languages is the challenge in their approach.

V. Malhotra, R. Singh (2018) analyzed and proposed evaluation metrics tailored for Indian language translation addressing the shortcomings in standard metrics for diverse linguistic contexts and advocated for context-aware evaluation to capture linguistic nuances.

S. Sundararajan, K. Rajan (2019) applied NMT to address morphological challenges in Tamil translation, highlighting the potential of neural models in handling complex linguistic structures. They reported significantly improved translation fluency and accuracy and concluded that resource-intensive training for morphologically rich languages.

Kapoor, R. Mehta (2020)explored multimodal translation incorporating both text and images for Indian languages, demonstrating the potential of multimodal approaches in enhancing translation performance and shown improved translation quality by leveraging visual context but with Increased model complexity and data requirements.

SenthamizhSelvi S and Anitha R (2022) proposed a hybrid POS-Tagger algorithm for Tamil language to effectively find equivalent words of Tamil in English using a limited English-Tamil parallel corpus.

Himanshu et al (2020) implemented a NMT model with Bi-LSTM, Muti-Head Self Attention and Byte Pair Encoding (BPE) for English-Tamil and English-Malayalam pairs and reported BLUE scores of 9.67 and 25.36.

Translating nouns and other domain specific terms requires transliteration. Transliteration is the process of converting text or words from one script to another. It involves representing the characters of one writing system with characters of another system. The purpose of transliteration is to accurately convey the pronunciation of words. YashMadhani et al. (2023) provide a transliteration corpus, made publicly available for 20 Indian official languages.

4.2 Challenges in Machine Translation for Indian languages

4.2.1 Ambiguity

Ambiguity involves multiple meanings or interpretations of words and phrases, requiring systems to discern the intended sense based on context. Word embedding like Word2Vec and GloVe, can be used to capture semantic relationships and aid in disambiguation

4.2.2 Polysemy

Polysemy refers to the phenomenon where a single word has multiple related meanings. Identifying the correct sense of a polysemous word (Sense Ambiguity) is crucial for accurate translation, requiring systems to disambiguate between related meanings.

4.2.3 Limited Parallel Corpora

There is a scarcity of parallel corpora (pairs of sentences in source and target languages) for training. This lack of data makes it challenging to train accurate and robust translation models.

4.2.4 Word Ordering

The Indian regional languages unlike English have a free-order sentence forms that do not demand Subject-Verb-Object pattern, leading to difficulties in capturing the diverse context.

5. RESULTS

The choice between RBMT, SMT, and NMT depends on factors such as the availability of resources, the complexity of the translation task, and the specific linguistic characteristics of the languages involved. NMT has emerged as the dominant approach, leveraging its ability to automatically learn complex patterns and adapt to diverse translation scenario. Table 1 shows some f the results of Machine Translation works from the literature.

Ref	*Method*	*Lang Pair*	*BLEU*
Koehn et al. (2003)	SMT	SV-EN	34.58
Koehn and Monz (2006)	SMT	EN-FR EN-ES EN-DE	28.33 27.49 14.01
Koehn and Hieu Hoang. (2007)	SMT	EN-ES EN-CZ EN-DE	24.25 27.62 18.22
Sutskever et al. (2014)	NMT	EN-FR	34.8
Wu et al. (2016)	NMT	EN-FR EN-DE	40.35 26.3
Johnson et al. (2017)	Zero-Shot translation (NMT)	PR-ES	24.75
Vaswani et al. (2017)	NMT+Attn+ Transformer	EN-DE EN-FR	28.4 48
Himanshu et al (2020)	NMT+Attn+BPE	EN-ML EN-TA	25.36 9.67

Table 1: Performance of existing systems

6. CONCLUSION

This comprehensive literature survey reviewed the main approaches to machine translation spanning from RBMT to NMT era. The survey provided an in-depth analysis of recent advances in machine translation, with a particular focus on transfer learning and transformer models. The integration of multimodal information, combining text and visual inputs, represents an exciting avenue for exploration. The survey discussed the intricacies of translation in Indian languages. Identifying and addressing challenges in machine translation is crucial for the continued advancement of the field. This survey serves as a snapshot of the current state of the field, recognizing achievements, confronting challenges, and paving the way for a future where machine translation systems are accurate and efficient. The synthesis of insights from an extensive range of research provides a holistic understanding of the evolving nature of machine translation, laying the groundwork for future developments and improvements.

7. LIMITATIONS

Machine Translation (MT) has witnessed significant advancements over the years, with Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) emerging as prominent paradigms. However, each

approach comes with its set of limitations, shaping the landscape of MT research and development.

RBMT, while conceptually robust, faces challenges in handling language intricacies and idiosyncrasies. The rigid rule-based nature often struggles with capturing the dynamic and context-dependent nature of language, resulting in suboptimal translations, particularly for languages with complex syntax and semantics.

SMT leverages statistical models to infer translation patterns from large bilingual corpora. Despite its success in certain scenarios, SMT exhibits limitations in capturing long-range dependencies and contextual nuances. The reliance on statistical probabilities can lead to inaccuracies, especially when dealing with idiomatic expressions, rare phrases, or low-resource languages. Additionally, SMT models often struggle with handling morphologically rich languages, impacting translation quality.

NMT has shown remarkable achievements by employing neural networks to learn complex mappings between source and target languages. However, the demand for extensive computational resources during training and inference poses a barrier for resource-constrained environments. NMT systems are also prone to producing fluent but contextually incorrect translations, and they can exhibit sensitivity to training data biases, potentially perpetuating stereotypes and cultural biases in translations

While this survey has provided a comprehensive overview of existing approaches in machine translation, there are several limitations that should be acknowledged to guide future research efforts. The survey focused on various paradigms, including neural, statistical, and rule-based methods, but the aspects deserve further attention.

- Ethical considerations in MT
- Evaluation Metrics
- Domain Specific Translation

REFERENCES

- M. f. Alawneh, T. M. Sembok and M. Mohd, "Grammar-based and example-based techniques in machine translation from English to Arabic," (2013) 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M), Rabat, Morocco, 2013, pp. 1-6, doi: 10.1109/ICT4M.2013.6518910.
- Sinha, R. M. K., Sivaraman, K., Agrawal, A., Jain, R., Srivastava, R., & Jain, A. 1995 ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In 1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century (Vol. 2, pp. 1609-1614). IEEE.
- ArviHurskainen and Jörg Tiedemann. 2017. Rule-based Machine translation from English to Finnish. In Proceedings of the Second Conference on Machine Translation, pages 323–329
- Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Mercer, R. L., ... &Roossin, P. S. 1990. A statistical approach to language translation. Computational Linguistics, 16(2), 79-85.
- Koehn, P., Och, F. J., &Marcu, D. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 48-54).
- Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.
- Och, F. J., & Ney, H. 2004. The alignment template approach to statistical machine translation. Computational Linguistics, 30(4), 417-449.
- Koehn, P., & Knight, K. 2002. Learning a translation lexicon from monolingual corpora. In Proceedings of the ACL-02 workshop on unsupervised lexical acquisition-Volume 4 (pp. 9-16).
- Foster, G., & Kuhn, R. 2007. Mixture-Model Adaptation for SMT. In Proceedings of the Second Workshop on Statistical Machine Translation (pp. 128-135).
- Koehn, P., &Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. In Proceedings of the Workshop on Statistical Machine Translation (pp. 102-121).
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).
- Bahdanau, D., Cho, K., &Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, IlliaPolosukhin. 2017. Attention is All You Need. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) pp. 5998-6008
- K. Chen, R. Wang, M. Utiyama, E. Sumita and T. Zhao, 2019. "Neural Machine Translation With Sentence-Level Topic Context," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 1970-1984,

- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, ..., and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ...& Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5, 339-351.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1412–1421,
- Devlin, J., Chang, M.-W., Lee, K., &Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)
- J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 2019 "mBERT: Pretraining of Multilingual BERT," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 4171–4186
- Lu, J., Yang, J., Batra, D., & Parikh, D. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in neural information processing systems (pp. 13-23)
- A. Gupta and B. Patel 2020 "Neural Machine Translation for Indian Languages," IEEE Transactions on Neural Networks and Learning Systems
- S. Kumar and M. Verma2015 "Rule-Based Machine Translation for Hindi-English," IEEE Transactions on Audio, Speech, and Language Processing
- R. Menon and S. Nair 2018 "Statistical Machine Translation for South Indian Languages," IEEE Transactions on Pattern Analysis and Machine Intelligence
- N. Das and A. Ghosh 2017 "Hybrid Approach for English to Bengali Translation," IEEE Transactions on Computational Intelligence and AI in Games
- P. Joshi and R. Kapoor, 2019. "Deep Learning for Code-Switching Translation in Indian Languages," IEEE Transactions on Multimedia,.
- K. Rao and S. Desai, 2021 "Low-Resource Language Translation: A Case Study of Kannada," IEEE Transactions on Knowledge and Data Engineering
- M. Singh and A. Jain 2016 "Cross-Lingual Transfer Learning for Hindi Translation," IEEE Transactions on Neural Networks and Learning Systems
- V. Malhotra and R. Singh 2018, "Evaluation Metrics for Indian Language Translation," IEEE Transactions on Audio, Speech, and Language Processing.
- S. Sundararajan and K. Rajan2019. "Neural Machine Translation for Morphologically Rich Languages: A Case of Tamil," IEEE Transactions on Neural Networks and Learning Systems.
- A. Kapoor and R. Mehta 2020. "Multimodal Machine Translation for Indian Languages," IEEE Transactions on Image Processing
- SenthamizhSelvi S &Anitha R 2022."Bilingual Corpus-based Hybrid POS Tagger for Low Resource Tamil Language: A Statistical approach", Journal of Intelligent and Fuzzy Systems, vol. 41, no. 6, pp. 8329–8348
- YashMadhani, SushaneParthan, Priyanka Bedekar, GokulNc, RuchiKhapra, AnoopKunchukuttan, Pratyush Kumar, and MiteshKhapra. 2023. Aksharantar: Open Indic-language Transliteration datasets and models for the Next Billion Users. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 40–57, Singapore. Association for Computational Linguistics.
- HimanshuChoudhary, Shivansh Rao, and Rajesh Rohilla. 2020. Neural Machine Translation for Low-Resourced Indian Languages. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3610–3615, Marseille, France. European Language Resources Association.

Enhanced Version of K4 Keyboard for the visually challenged

Dr. V. Krishnamoorthy

ABSTRACT

The following changes, simplifications and enhancements have been made in the second version of K4 Keyboard, a multilingual keyboard for the touch screen smart phones, for the visually challenged. One major change is the removal of the requirement that, for a character or a vowel extension, the swipe 'returns to the starting point'. This means that when the information content of a stroke is over, there is no need to return to the starting point, but just continue with the next letter. This saves about half of the input time. The next major change is that the elimination of remembering the swipes for the many commands and symbols. Now, just swipe the names of these. We have also simplified the way the dynamic abbreviations are swiped. Currently we are concentrating on English and Tamil. In English, we have introduced an automatic error correction of one change in any one letter after the fourth letter. The concept of gap has been replaced by 'one line per character'. This reduces the options considerably. In Version 1, Tamil words with a limited number of extensions only were recognized from a swipe. In Version 2, we can get any word with any number of extensions with one swipe. We have also introduced a provision to get the correct la, ra, na.

Dr. V. Krishnamoorthy
Former Professor, Anna University.
Learnfun Systems, prof.vkrish@gmail.com

INTRODUCTION

Currently the visually challenged are using the smart phones with the touch screen keyboards to input a text as follows. They touch a point on the keyboard, and keep the finger on the phone. The system reads out the letter touched. If it is the intended letter, a double tap confirms that letter. Otherwise, the finger is moved towards the required letter. When the finger enters a letter that letter is read out. The finger is taken back when the required letter is reached, and a double tap confirms that letter. Any text is input letter by letter in this way. Word prediction is available, but the words shown have to be tapped to find out what that word is. This may not be efficient always.

The normal swipe method used by normal people for fast input is not available for the visually challenged.

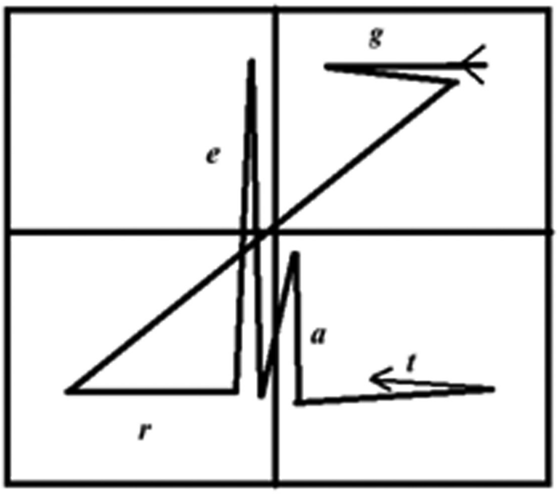
To speed up the input of text by the visually challenged, we developed a keyboard called K4 keyboard. Its first version was out in June 2019. It had only 4 keys. Each square key had the side of half the width of the phone. Hence there is no chance for selecting a wrong key. The inputs of letters were by small swipes. A full word can also be input with one swipe. We created a method in which the letters had to be swiped, one after the other, continuously. We also introduced methods to swipe a whole phrase. Editing functions, spell checking and search and replace were included using specific swipes. Numbers could also be input with just one swipe. Symbols and emojis could be got using a method we called dynamic abbreviation. It catered to about 75 languages. We had put as many things as we could think of. Our motivation at that time was that we can do so much with the new keyboard.

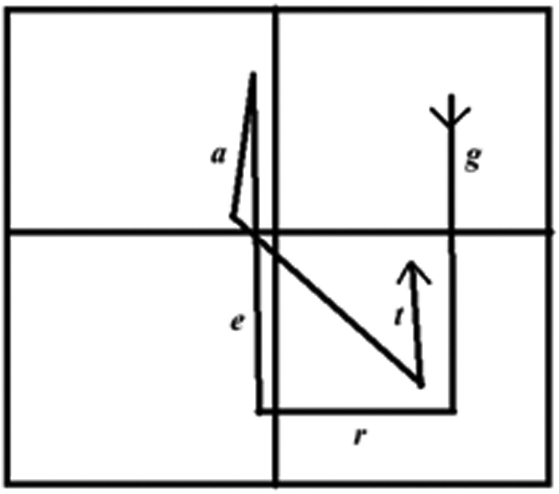
We did not do any marketing. We assumed it will be propagated by word of mouth. But it didn't happen. After some time we started to think of the reasons for this. We came up with two reasons. One is that the explanation provided in the help was difficult to understand. Another thing is that we had put too many things in the software. We wanted to rectify these. We started thinking about simplifying the way the swiping was explained. This led us to think of simplifying the swiping process itself. This lead to quite a large number of iterations and changes. Finally we could redesign the swipe method, and that reduced the swiping time

considerably! Now the saved input time is around a whopping 40%.

In Tamil and other South Indian languages, a noun or verb can be transformed by adding many parts like the vaetrumai vurubu. There can be thousands of such variations for a word. Any real software for Tamil had to take care of all these modifications. In our first version of K4 keyboard, we could not do this, since our algorithms at that time could not do that within a reasonable amount of time. Note that a word has to be predicted within about a second, to make the keyboard usable. In the current second version of K4 keyboard we have overcome this difficulty. Now all the Tamil words are taken care of. The software is ready for field trail. We have made more than half a dozen changes from the first version. This paper explains the changes made.

Change 1: Reduce the swipe length





Let us explain using an example. Consider the word 'great'. In the first version the swipe for that was as shown in the first figure above. It starts from key 2, in direction 3, that is move left from the top right key. The letters are shown near the lines which represent them. Note that though we have shown the required letter near a line, that line may represent more than one letter. It is the business of the software to predict the required letter. After the letters g, e, and a, note that line backtracks. It has 10 lines. The second figure above shows the way this word is swiped in the next version. The letter starts in the downward direction. This is because, we have changed the paces of the alphabets, to be consistent with their placement in the 12 key phone keyboard. In this a, b, c start in direction 2, that is going to the right from top left key. There are no backtracking lines in the new method. In this example it has only 6 lines, and the number of lines has reduced by 40%. This great saving reflects in the input time. This type of change has been implemented for the swiping of numbers also.

Change 2: Commands using dynamic abbreviation

In the previous version, all the commands and many symbols were given specific swipes. One had to remember the swipes to use them. In the new version, these are got by dynamic abbreviation. This is a method we had invented in the previous version, for fast input of phrases, mathematical symbols, and emojis. In this method, a phrase is got by giving the first few letters from some words of the phrase. The words chosen, and the number of letters chosen are not fixed, and can vary every time. A symbol is got using the dynamic abbreviation of their names. In the same way now a command is got from their names. The difference is that the system will predict a command. If that is the intended command, it has to be confirmed by a small swipe. This new method eliminates remembering the numerous swipes for commands.

Change 3: New method for dynamic abbreviation

The way a dynamic abbreviation is entered has been changed. In the new method, first we have to specify which data base is to be used and whether the abbreviation uses a single letter from a word or multiple letters from a word. The keys 1 to 4 represent the commands, emojis, phrases and symbols in that order. A single tap to start with means that only one letter per word will be chosen. A double tap will indicate that multiple letters are chosen from a word. A single or double tap on any of the four keys specifies the data

base and the number of letters chosen. After this tapping the letters are input with just one swipe.

Change 4: Automatic one error correction in English

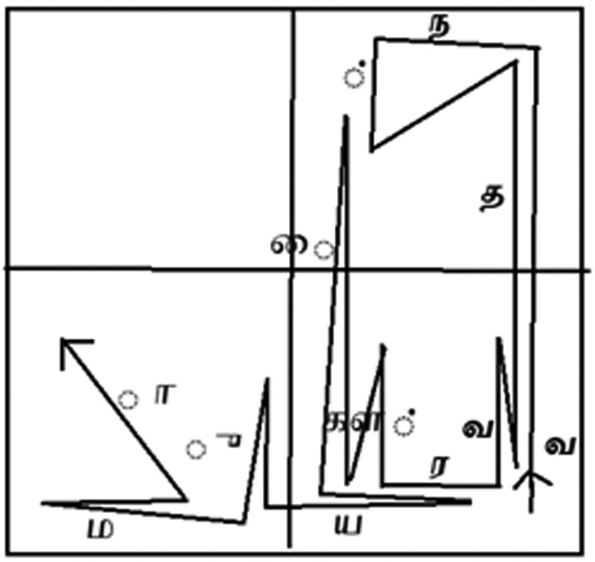
For English, we have included an automatic error correction. When the swipe does not give a word, it will check for one error in any letter after the fourth letter and get the possible words.

Change 5: Diagonal to represent any letter

In the previous version, a diagonal and back was used to indicate any number of letters. This resulted in the prediction of a large number of words in some cases. This is now modified as follows. A diagonal represents exactly one letter. A sequence of diagonals represents that many letters. The letters can be anything. This fixes the number of letters in a word. This limits the number of predictions very much. This method speeds up the input og long words. This method is to be used after the fourth latter only.

Change 6: All Tamil words can be swiped now

In the previous version, we could predict only a limited number of extensions of a word in Tamil. In the new version, any word in Tamil, with any number of extensions, will be predicted within about a second. We have successfully modified the spell check, and also made it faster. Now this new intelligent use of spell check can be used for the other south Indian languages, Telugu, Kannada and Malayalam also. The power of this method can be seen using an example.



Consider the word வந்தவர்களையுமா. It has 7 parts. The step by step formation of this word can be seen as follows. வா ந்த வர் கள் ஐ யும் ஆ. The Unicode characters for this word are வ ந ் த வ ர ் க ள ை ய ு ம ா. There are 14 Unicode characters in this word. The swipe for this word has got just 19 lines! This word gets predicted in about one second. As indicated in the case of English example, only the required letters are shown near the corresponding lines. These lines usually represent many letters. The software predicts the required letters. Note that all the thousands of extended words cannot be kept in a static data base. All these variations have to be generated dynamically. Here one line can represent many letters. Choosing the correct letter from a list, and also checking from a dynamic dictionary was a very difficult task. We have achieved that and also kept the time of execution quite acceptable.

Change 7: Error correction for la, ra, na letters

Sometimes one gets a doubt about which la or ra or na is to be used in a word. We have some help in this regard. Just after swiping the doubtful la or ra or na letter, a small diagonal and back, tells the system to check for the other letters also. When more than one letter give meaningful words, some help regarding the meaning of these words are given.

CONCLUSION

The method of swiping has been redefined. The major change is the reduction of lines needed to swipe a word. This has resulted in reducing the input time considerably. All the Tamil words can now be got with a swipe. The help is rewritten and simplified. All the above changes have been made after many many iterations. The system is ready for field trial for English and Tamil. The other languages will be introduced later one by one, once the software is stabilized for English and Tamil. Because of these simplifications, we hope that this version of our keyboard for the visually challenged will be welcomed by the community.

Legal-MT: Building English-Tamil Neural Machine Translation System for Judiciary Domain

Ramakrishna Appicharla, Asif Ekbal

ABSTRACT

A large amount of legal content is produced daily in a multilingual country like India. Due to this, there is a need for domain-specific machine translation (MT) systems that can translate legal data from one language to another. Manual translation of legal data is challenging and time-consuming. The amount of time and effort can be reduced by developing an MT system to aid the human translator during the translation. However, building a high quality MT system is difficult due to the unavailability of a parallel corpus. In this work, we create a neural machine translation system (NMT) to translate the judiciary context from English to Tamil. We train our model on publicly available English-Tamil judiciary parallel corpus extracted from various court judgments and court orders of Indian courts. Our model achieved a 42.3 BLEU score on a held out test set of 8,729 sentences. We also evaluated our model on WAT‘21 and Flores-200 test sets, and our model achieved BLEU scores of 8.0 and 15.8, respectively. Finally, we deploy our MT system for public use, and it is accessible at:

http://hemat.in.ngrok.io/

Ramakrishna Appicharla, Asif Ekbal

Department of CSE, Indian Institute of Technology Patna, India.
{appicharla_2021cs01, asif}@iitp.ac.in

1. INTRODUCTION

India is a multilingual country with significant linguistic and cultural diversities. People speak many different languages, and there is a growing need to make useful information available in the vernacular languages. Many legal documents are produced daily in an extensive, highly populated country like India. The legal domain has continuous publishing cycles, and the growing demand for multilingual information access requires these documents to be translated into a language that the end user understands most. In that respect, the burden of granting access to this information falls on the government, which must manage a continuous cycle of document translation needs. A certified translator will take a long time to translate documents related to legal Proceedings, FIRs (First Information Reports), petitions, judgments, etc. The number of profes- sional translators in this field is limited, which takes 8-10 hours for ten pages. Therefore, this will in- troduce a significant delay in the overall judgment comprehending process, equivalent to “Justice de- layed is justice denied.” An automated machine translation system will help in this process. The documents could be translated using an automated translation system, which can then be post-edited by human translators.

Machine translation (MT) has seen tremendous progress with the advancement of training large neural networks (Krizhevsky et al., 2012). Neural Machine Translation (NMT) (Sutskever et al., 2014; Bahdanau et al., 2014; Gehring et al., 2017) which uses deep neural networks to train an end- to-end model to achieve significant improvements over the Statistical Machine Translation (SMT) (Bojar et al., 2016) based approaches. Recently, transformer-based (Vaswani et al., 2017) encoder- decoder models achieve better results than the re- current neural networks (RNN)-based (Sutskever et al., 2014; Bahdanau et al., 2014) models, making them de facto in developing any MT applica- tion. As the neural networks can be trained end-to- end, various approaches have been proposed such as multilingual NMT multilingual NMT (John- son et al., 2017; Liu et al., 2020), unsupervised NMT (Artetxe et al., 2018; Lample et al., 2018), and document-level NMT (Voita et al., 2018; Huo et al., 2020) to improve the translation performance. However, the NMT-based system achieves near human-level performance

(Hassan et al., 2018) in languages like German, French, English, etc., these systems suffer from issues such as data scarcity, noise in the training data, and poor domain generalization (Koehn and Knowles, 2017). Such problems are complex when developing a high-quality, domain-specific MT system for low-resource lan- guages such as English-Tamil. The development of a legal NMT system can reduce the human efforts to translate judiciary documents, but the availability of such data is limited. In this work, we aim to develop a high-quality NMT system to translate judiciary data from English into Tamil by extracting available English-Tamil judiciary parallel corpus. We report BLEU (Papineni et al., 2002), ChrF (Popovic´, 2015), and TER (Snover et al., 2006) scores on a held-out judiciary test set along with the results from WAT'21 (Nakazawa et al., 2021) and Flores-200 (Costa-jussà et al., 2022) test sets. We also deploy our system[1] for public use.

2. RELATED WORK

Sutskever et al. (2014) proposed the encoder- decoder framework, which uses RNNs for the task of translation. The performance of the encoder- decoder networks is further enhanced by the introduction of attention (Bahdanau et al., 2014). Gehring et al. (2017) proposed the convolutional neural network (CNN)-based NMT system, which further improved the training times of NMT systems. Vaswani et al. (2017) introduced the trans- former networks, which are the current state-of-the-art neural network-based approach for many tasks, including machine translation. Similarly, Johnson et al. (2017) proposed an approach to train a multilingual NMT model with a single encoder-decoder network. Several techniques, such as byte- pair encoding (Sennrich et al., 2016b) and back- translation (Sennrich et al., 2016a), have become the standard practices for training high-quality NMT systems.

There have been some efforts to develop MT systems for the judiciary domain. Farzindar and Lapalme (2009) developed an SMT system to translate court judgments between English and French. Similarly, Martínez-Domínguez et al. (2020) developed NMT systems for the Swiss legal do- main between French-German, French-English, Italian-French, and German-Italian languages. Ive et al. (2020) created a high-quality legal translation dataset for English-Dutch, English-French, and English-Portuguese language pairs. However, there are few efforts to develop judiciary MT systems for Indic languages. Mahapatra et al. (2023) created a high-quality judiciary parallel corpus between English and nine other Indic languages and reported the performance of available MT systems on the prepared corpora. In this work, we develop an NMT system to translate judiciary data from English to Tamil. We deploy our system for public use.

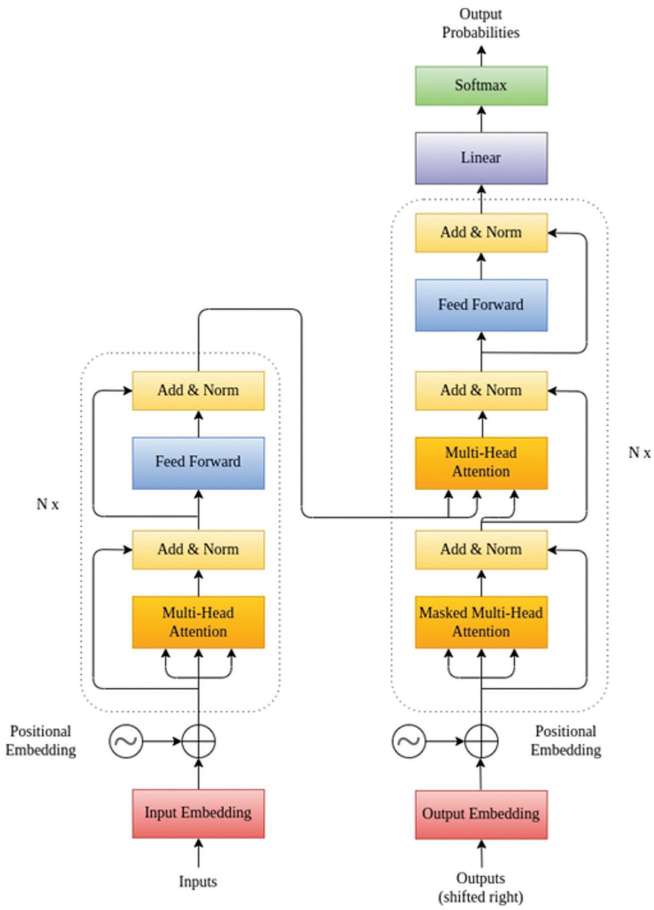


Figure 1: The overview of the Transformer architecture. Here, N denotes the number of layers. The output is shifted right to make the decoder predict the current output token based on previously generated tokens.

3. METHODOLOGY

Our NMT system is based on transformer (Vaswani et al., 2017) architecture (cf. Figure 1). The trans- former architecture follows the encoder-decoder (Sutskever et al., 2014) approach, where two separate networks (encoder and decoder) are used to translate the source sentence into its equivalent target translation. Specifically, the input (source sequence) is encoded by the encoder network into an intermediate representation, which is then used by the decoder to produce the output (target sequence). The multi-head attention (MHA) layer is the main component of the transformer network. The MHA layer computes attention in parallel by attending to different parts of a given sequence. The parallelism is controlled by specifying a number of heads during the attention computation. Query (Q) and Key-Value (K − V) pairs are required in order to calculate the MHA, and the MHA is calculated as:

1. http://hemat.in.ngrok.io/

$$Attention\,(Q, K, V) = Softmax\left(\frac{QKT}{\sqrt{dk}}\right)\, V \qquad (1)$$

$$MultiHead\,(Q, K, V)\; Concat\,(head_1, \ldots, head_h)\; W^O \quad (2)$$

$$Head_i = Attention\,(QW_i^Q, KW_i^K, VW_i^V) \qquad (3)$$

The MHA sub-layer receives the encoder's input and uses it to compute self-attention. The output from the MHA sub-layer is fed to the position wise feed forward sub-layer, which is a two layer feed forward network with the rectified linear unit (ReLU) (Glorot et al., 2011) activation. The output from each sub-layer is normalized by applying the layer normalization (Ba et al., 2016) with residual connection (He et al., 2016). The resulting output is fed to subsequent layers (typically, 6-layered stacks are used), and the output from the last layer is considered an encoded representation of the input source sentence.

Like the encoder network, the decoder network also consists of a stack of 6-layers. Along with MHA and position-wise feed-forward layers, each decoder layer contains an additional masked MHA, which computes the self-attention between target tokens. All future tokens concerning the current target token are masked to prevent the model from remembering the future positions in the target sequence. A linear layer with softmax activation estimates the likelihood of the subsequent word in the sequence. The input source and target sequences are sent to the encoder and decoder layers after being token and position-wise embedded. Given a sentence pair (x, y), the network is trained to maximize the log-likelihood as:

$$p(\mathbf{y}|\mathbf{x};\theta) = \prod_{m=1} p(y_t|x, y_{<m}) \qquad (4)$$

$y_{<m}$ is the partial target sequence generated till time step m. The trained parameters of the model $(\hat{\theta})$ are used to generate the most likely translation $(\hat{\mathbf{y}})$ based on maximum-a-posteriori approximation (MAP) as:

$$\hat{\mathbf{y}} = \underset{\mathbf{y}}{argmax}\, p(\mathbf{y}\,\mathbf{x};\hat{\theta}) \qquad (5)$$

4. EXPERIMENTAL SETUP

4.1 Data Statistics

We build an NMT model to translate English judicial content into Tamil. We use judicial domain Parallel corpus from Anuvaad project[2] to train the model. The corpus consists of judicial data extracted from court judgments and court orders of Indian courts. The corpus contains a total of 816,729 parallel English-Tamil sentence pairs. We train the model with 800,000 sentence pairs, and 8,000 sentence pairs are used as validation sets. We evaluate the performance of the trained model on

Three different test sets viz.a held out test set of 8,729 sentences from the Anuvaad corpus, WAT'21 (Nakazawa et al., 2021) test set[3] of size 2,390 sentences and Flores-200 (Costa-jussà et al., 2022) devtest set[4] of size 1,012 sentences.

4.2. NMT Model Setup

We train a transformer (Vaswani et al., 2017) based NMT model. We employ a six-layered encoder-decoder stack with eight attention heads. The dropout rate is set to 0.1 (Srivastava et al., 2014), while the embedding size and hidden sizes are [5]12. The position-wise feed-forward layer consists of 2048 cells. The model is optimized with Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 2 and 8,000 warm-up steps. The learning rate during the training is updated based on the Noam (Vaswani et al., 2017) learning rate scheduler. We create two separate sub-word vocabularies of size 35,000 based on the unigram language model (Kudo, 2018), using Sentence Piece (Kudo and Richardson, 2018) implementation. We use OpenNMT toolkit (Klein et al., 2017)5 to train our model with a token-level batch size of 2048 tokens. Checkpoints are created for every 10,000 steps, and all checkpoints created during the training are aver- aged and considered as best model parameters. We perform a beam search during the inference with a beam width of 5 and no length penalty.

5. RESULTS AND WEB INTERFACE

5.1. Main Results

We report BLEU (Papineni et al., 2002)[6], ChrF (Popovic´, 2015)[7], and TER (Snover et al., 2006)8

Test set	BLEU (↑)	ChrF2 (↑)	TER (↓)
Law	42.3	78.2	65.5
WAT'21	8.0	50.1	89.0
Flores	15.8	57.7	77.8

Table 1: BLEU, ChrF2, and TER scores of the trained model on different test sets. Higher BLEU, ChrF2, and lower TER scores are better.

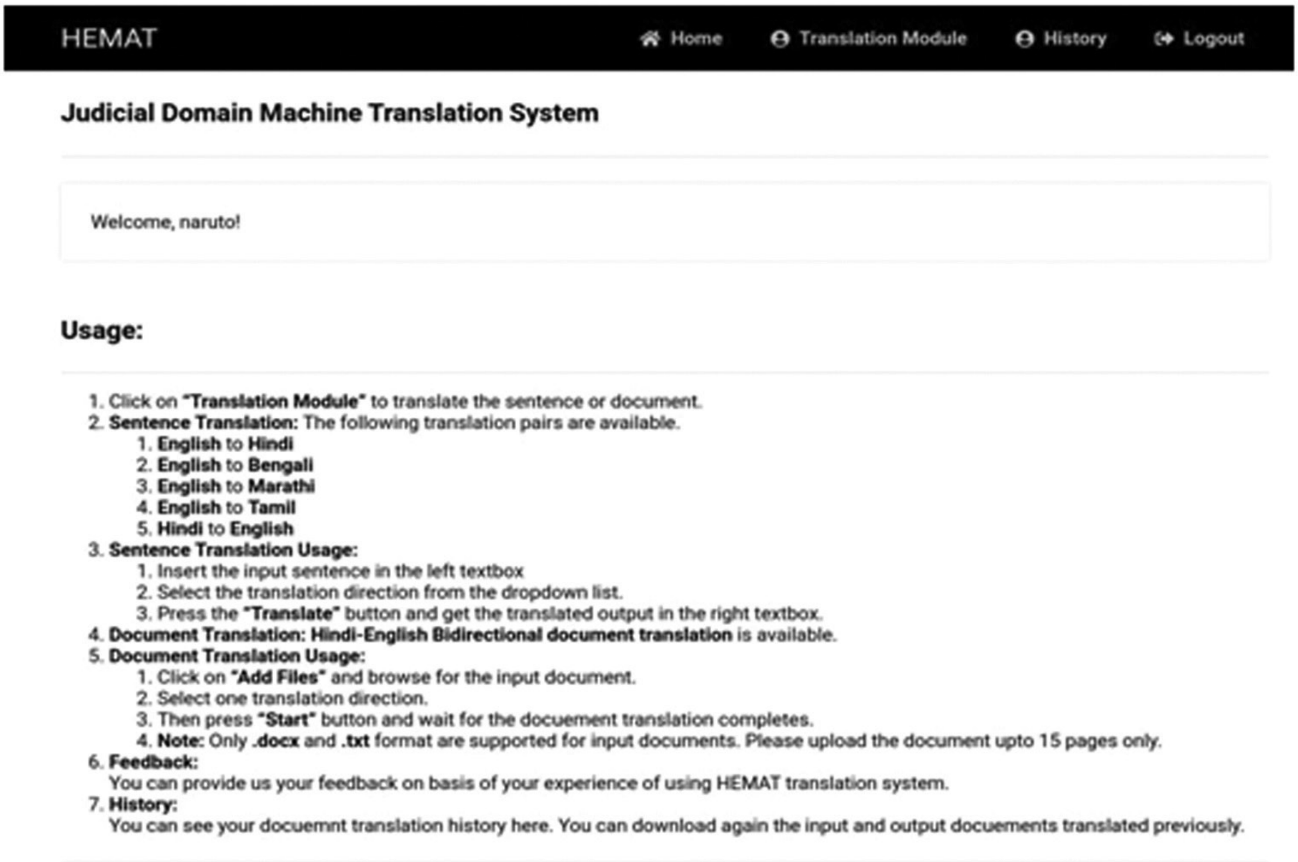
Scores calculated with sacreBLEU (Post, 2018)[9] toolkit. Table 1 shows the results of the trained English-to-Tamil model on different test sets. The trained model performs well on an in-domain (law) test set. However,

the model performs poorly on test sets from general domains. This is a well- known problem in NMT (Koehn and Knowles, 2017) where the performance of any domain- specific NMT system will degrade when tested with other domain data. Although the model's performance is lower for general domain data, the model can translate judiciary domain data. Our main goal is to reduce the human effort required to translate judiciary documents via the NMT system rather than an automated NMT system, which removes the human in the loop. The main reason for such a design is that judiciary documents contain sensitive content and human intervention is needed to verify the correctness and authenticity of the translation.

5.2. Web Interface

We deploy our system[10] for public use. Initially, the system is developed to translate between English. We extended the system to support other language pairs (such as English-to-Tamil, English-to- Marathi, and English-to-Bengali). Figure 2 shows the sample screenshots of the web interface to access the trained model. Users should register to use the system. After the registration/login, the user is presented with the Home page (cf. Figure

(a) in 2) where the instructions to the system are given. Currently, we support only sentence-level translation for English-to-Tamil direction, and the document translation support is only enabled for English-Hindi (bidirectional) direction. We plan to extend the document translation support for other language pairs.



(a)

9. https://github.com/mjpost/sacrebleu

10. http://hemat.in.ngrok.io/



(b)

Figure 2: Screenshots of the web interface. (a). Home page, (b). Sample sentence translation.

6. CONCLUSION AND FUTURE WORK

This paper describes developing a neural machine translation system for English-to-Tamil language direction to translate judiciary data. We train the model on the project Anuvaad judiciary corpus and report the model's performance in terms of BLEU, ChrF, and TER scores. We observe that the model can perform well on the judicial domain test set. However, the performance is lower on the general

Domain test sets. We deploy our model for public use and can be accessed at: http://hemat.in. ngrok.io/.

We plan to extend the current system to translate documents and produce the output in the given document format. We also plan to improve the performance of the model by introducing recent techniques such as multilingual training (Johnson et al., 2017), document-level translation (Voita et al., 2018; Kim et al., 2019; Voita et al., 2019) approaches.

Acknowledgements

Authors gratefully acknowledge the support from "NLTM: VIDYAAPATI" project, Sponsored by Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India.

REFERENCES

- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In International Conference on Learning Representations.
- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hin-ton. 2016. Layer normalization. arXiv preprint arXiv:1607.06450.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
- Ondˇrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, An- tonio Jimeno Yepes, Philipp Koehn, Varvara Lo- gacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Pa- pers, pages 131–198, Berlin, Germany. Association for Computational Linguistics.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No

language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

- Atefeh Farzindar and Guy Lapalme. 2009. Machine translation of legal information and its evaluation. In Advances in Artificial Intelligence: 22nd Canadian Conference on Artificial Intelligence, Canadian AI 2009 Kelowna, Canada, May 25-27, 2009 Proceedings 22, pages 64–73. Springer.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In International conference on machine learning, pages 1243–1252. PMLR.
- Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In Pro- ceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of Proceedings of Machine Learning Research, pages 315–323, Fort Lauderdale, FL, USA. PMLR.
- Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic Chinese to English news translation. arXiv preprint arXiv:1803.05567.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.
- Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural ma- chine translation. In Proceedings of the Fifth Conference on Machine Translation, pages 604–616, Online. Association for Computational Linguistics.
- Julia Ive, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura, and Maxim Khalilov. 2020. A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality? In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3692–3697, Marseille, France. European Language Resources Association.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: En- abling zero-shot translation. Transactions of the As- sociation for Computational Linguistics, 5:339–351.
- Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 24–34, Hong Kong, China. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
- Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senel- lart, and Alexander Rush. 2017. OpenNMT: Open- source toolkit for neural machine translation. In Pro- ceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. Association for Compu- tational Linguistics.
- Philipp Koehn and Rebecca Knowles. 2017. Six chal- lenges for neural machine translation. In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computa- tional Linguistics.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hin- ton. 2012. Imagenet classification with deep convo- lutional neural networks. Advances in neural infor- mation processing systems, 25.
- Taku Kudo. 2018. Subword regularization: Improv- ing neural network translation models with multiple subword candidates. In Proceedings of the 56th An- nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 66–75, Melbourne, Australia. Association for Computational Linguistics.
- Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tok- enizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase- based & neural unsupervised machine translation. In Proceedings of the 2018 Conference on Empiri- cal Methods in Natural Language Processing, pages 5039–5049, Brussels, Belgium. Association for Com- putational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre- training for neural machine translation. Transac- tions of the Association for Computational Linguis- tics, 8:726–742.
- Sayan Mahapatra, Debtanu Datta, Shubham Soni, Adri- jit Goswami, and Saptarshi Ghosh. 2023. Improving access to justice for the indian population: A bench- mark for evaluating translation of legal text to indian languages. arXiv preprint arXiv:2310.09765.
- Rubén Martínez-Domínguez, Mat¯ıss Rikters, Artu¯ rs Vasil, evskis, Ma¯rcis Pinnis, and Paula Reichenberg. 2020. Customized neural machine translation sys- tems for the Swiss legal domain. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), pages 217–223, Virtual. Association for Machine Translation in the Americas.
- Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondˇrej Bojar, Chenhui Chu,

Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In Proceedings of the 8th Workshop on Asian Trans- lation (WAT2021), pages 1–45, Online. Association for Computational Linguistics.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei- Jing Zhu. 2002. Bleu: a method for automatic evalu- ation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Compu- tational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popovic´. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186– 191, Brussels, Belgium. Association for Computa- tional Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Lin- guistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In Proceedings of the 54th An- nual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715– 1725, Berlin, Germany. Association for Computa- tional Linguistics.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of trans- lation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Se- quence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information pro- cessing systems, pages 5998–6008.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In Proceedings of the 57th An- nual Meeting of the Association for Computational Linguistics, pages 1198–1212, Florence, Italy. Asso- ciation for Computational Linguistics.
- Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine trans- lation learns anaphora resolution. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPLICATIONS FOR TAMIL COMPUTING

Smart Phone based Tamil Language Interfaced Digital Solutions for Stray Animal Welfare in Tamil Nadu: 'Find a Stray Animal (FISA)' - Stray Tracking App

Supraja Vaidhyanathan, Haripriya Chandrasekhar, Praveen Kumar Kannan, Vijay Jeyakumar, P. Selvaraj

ABSTRACT

"Find a Stray" is a proposed mobile application addressing the challenges of stray animal management, notably encompassing cows on Indian roads. Beyond dogs and cats, the app recognizes the prevalent issue of stray cows, particularly in India, where cultural and religious sentiments grant them free movement on roads. These stray cows contribute to traffic congestion and safety hazards, necessitating a comprehensive solution. The app harnesses smartphone technology to allow users to report and assist not only dogs and cats but also stray cows. By capturing images, utilizing geo-tracking and categorizing animals, the app connects users to nearby animal welfare organizations. Furthermore, "Find a Stray" engages users in community efforts by incorporating crowdfunding and resource donation campaigns. This multifaceted approach aims to address the complex cultural, religious, and urban development aspects associated with managing stray cows and other animals on Indian roads while fostering support for animal welfare organizations.

Supraja Vaidhyanathan, Haripriya Chandrasekhar, Praveen Kumar Kannan, Vijay Jeyakumar
Department of Biomedical Engineering,
Sri Sivasubramania Nadar College of Engineering, Kalavakkam.

&

P. Selvaraj
Lead PI, ICAR-NASF Artificial Intelligence & IoT Smart Vet Project,
Department of Veterinary Medicine, College of Veterinary Medicine, Chennai and
Professor and Head, Veterinary University Outpatient Hospital,
Tamil Nadu University of Veterinary and Veterinary Sciences, Madhavaram, Chennai.

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

In the heart of every community lies a concern for the welfare, safety, and health of its inhabitants, including the often overlooked but equally vital members – stray dogs and cats. In the absence of dedicated care, these remarkable beings navigate a challenging path, facing the risk of injury, accidents, and an uncertain fate. Recognizing the inadequacies of traditional methods in addressing this issue, we proudly introduce the Stray Dog Tracking App – a mobile solution designed to revolutionize stray animal management and contribute to their well-being.

At the core of our initiative lies a commitment to empowering communities to actively engage in the identification, tracking, and assistance of stray dogs. The app boasts a user-friendly interface that facilitates seamless navigation, making it accessible to individuals from all walks of life. Real-time GPS tracking and mapping features harness cutting-edge technology to offer precise monitoring of stray dog movements, providing a valuable tool for efficient and accurate identification.

The app does not stop at tracking; it invites users to become advocates for change.

Through an intuitive in-app reporting system, users can share detailed sightings of stray dogs, complete with photos and descriptions. This user-generated data becomes a powerful resource, streamlining the rescue and care process for these vulnerable animals.

Our commitment extends beyond technology. By establishing a robust connection between users and local animal shelters and rescue organizations, the app acts as a catalyst for collaborative efforts. This network facilitates the safe capture and care of stray dogs, turning the community into a collective force for good.

But our vision transcends the practical aspects of tracking and rescue. The Stray Dog Tracking App aims to be a platform for education and awareness. By leveraging its reach, we aspire to shed light on the challenges faced by stray dogs, fostering empathy and understanding within communities. Through a

united front, we encourage community involvement in advocating for the welfare of these fantastic beings.

Beyond tracking and awareness, the app strives to be a catalyst for change by actively seeking foster or adoption homes for stray dogs. In doing so, it not only addresses immediate concerns but also provides a tangible pathway for these animals to find loving homes, breaking the cycle of uncertainty.

1.2 Literature review

The literature on stray animal population control, legal treatment of stray animals, beliefs about feeding strays, and the impact of stray animal-related road traffic accidents underscores several key issues.

Firstly, these studies emphasize the significant global concern regarding the overpopulation of stray animals. Stray animals pose various challenges, including public health risks, zoonotic disease transmission, road accidents, and ethical considerations regarding their treatment and control methods. However, there's limited empirical research focusing specifically on India's context, where the issue of stray animal population control is notably prominent.

Secondly, legal perspectives and legislative measures concerning stray animals are highlighted. The study discussing changes in Russian legislation indicates the importance of legal frameworks in managing stray animals. However, there's a lack of comprehensive research exploring the legal implications and measures for stray animal management within the Indian legislative context.

The third study delves into human beliefs regarding feeding stray animals. It sheds light on the complexity of human-animal interaction, emphasizing the need for tailored interventions considering diverse beliefs and backgrounds. However, such in-depth investigations into societal attitudes towards feeding stray animals are infrequently conducted in India, despite its relevance to local practices and interventions.

Lastly, the impact of stray animal-related road traffic accidents is a significant concern addressed in the literature. The study analyzing injury patterns resulting from animal-vehicle collisions emphasizes the need for preventive measures and public awareness. This raises concerns about the lack of extensive studies specifically focusing on the patterns, implications, and preventive strategies for stray animal-related road accidents within the Indian context.

Overall, while these studies offer valuable insights into stray animal management, ethical concerns, legal frameworks, societal attitudes, and public health implications, there is a notable scarcity of comprehensive studies and data specifically focusing on India's unique challenges and circumstances regarding stray animal populations and related issues. Further empirical research in India could significantly contribute to understanding and addressing the multifaceted challenges posed by stray animals in the country.

2. APP DESIGN

The app offers a robust set of features to assist users in reporting and managing stray animals effectively. Upon registration or login through various credentials including social media, users can access their previous history of reported animals, keeping track of their status. The app seamlessly toggles between English and Tamil languages via a language icon, accommodating users' language preferences. Leveraging regional language processing, users can input information in Tamil, ensuring inclusivity and ease of use.

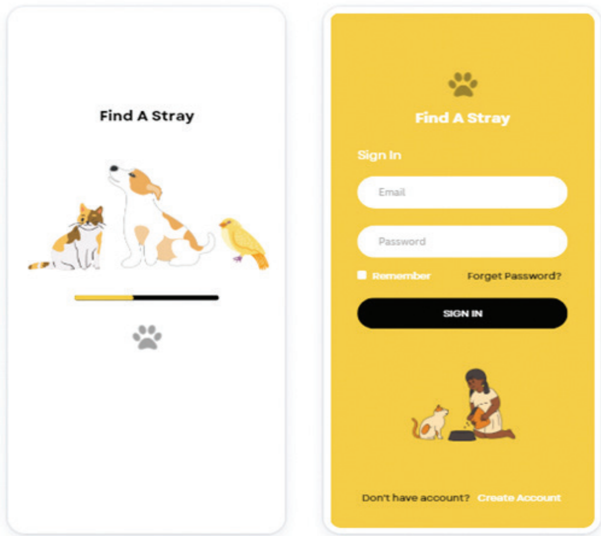


Figure 1: Front end of the app: registration phase

Utilizing image recognition and geotagging technology, the app employs a trained model to process uploaded images of strays, extracting relevant data, and connecting users with organizations like Blue Cross or vet hospitals for assistance. Additionally, pet owners can engage in telemedicine services for their pets, scheduling virtual appointments or calls with veterinarians. Secure payment gateways are integrated for seamless transactions related to telemedicine or other in-app services.

The app's functionality extends to providing updates on the status of reported animals, allowing users to track the well-being of animals taken for care or treatment. Moreover, the app facilitates connections with foster homes for those interested in adopting or fostering stray animals. Users reporting stray animals can input crucial details such as location and identifying features to ensure accurate identification and prompt assistance for the animals in need.

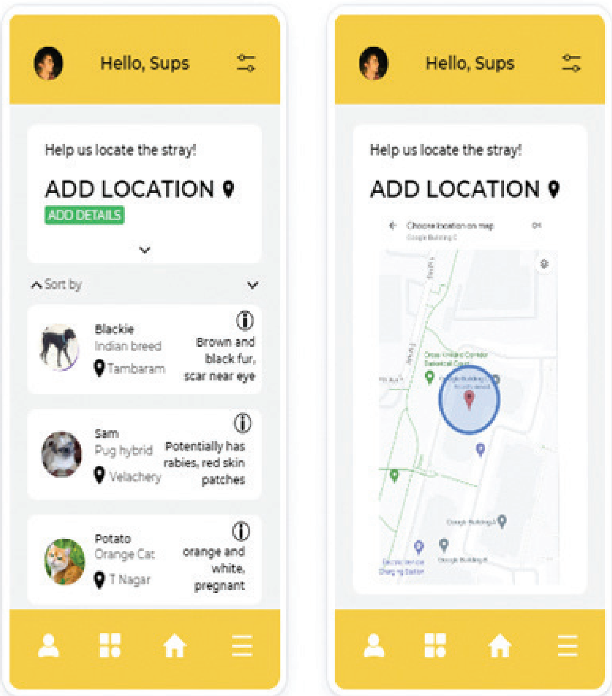


Figure 2:
Front end of the app: Location identification phase

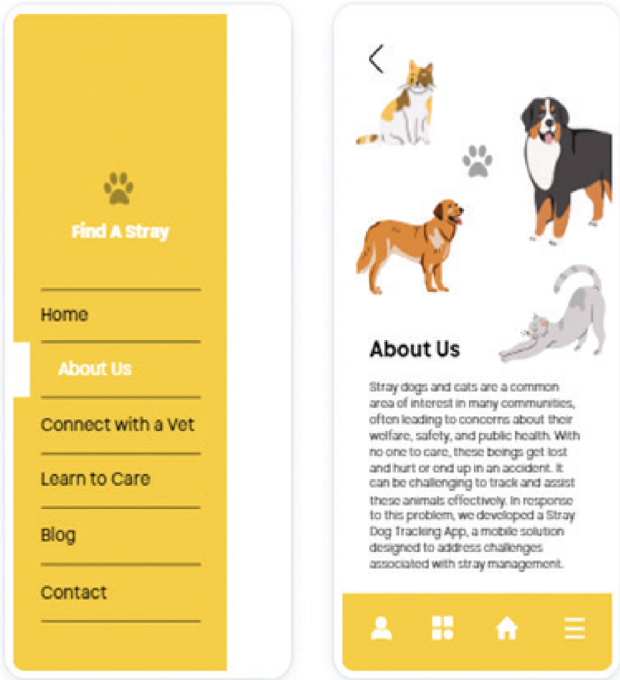


Figure 4:
Front end of the app: Content of the app

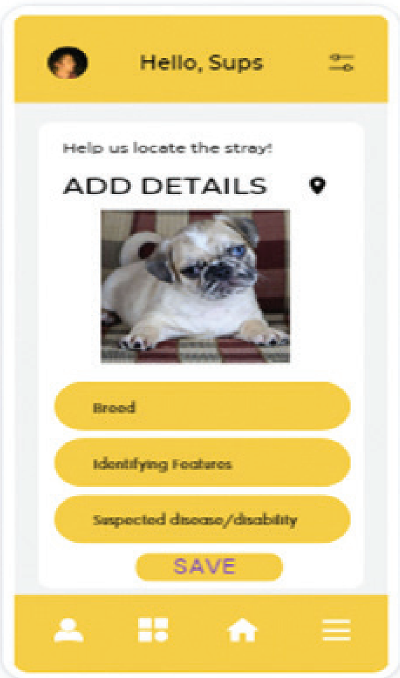


Figure 3:
Front end of the app: Stray detail uploading phase

3. ETHICAL CONSIDERATIONS

Developing and deploying the Stray Dog Tracking App hinges on our steadfast dedication to upholding ethical standards that prioritize the welfare, dignity, and rights of both human and animal stakeholders involved. Our commitment to ethical practices forms the foundation of the app's design and implementation, ensuring the well-being of stray dogs and the community members engaging with the application.

First and foremost, our ethical considerations revolve around the welfare of stray dogs. The app's primary objective is to enhance the lives of these animals. We prioritize humane treatment and ensure that the app's tracking and reporting functionalities aim to aid in the rescue and care of stray dogs without causing harm, distress, or disruption to their natural behaviours. Stray Dog Tracking App's focus remains steadfast on improving the animals' conditions while respecting their autonomy and well-being.

Respecting user privacy is another core ethical principle embedded in the app's development. Robust data protection measures are implemented to safeguard users' personal information. All data collected, whether through GPS tracking or user-generated reports, is handled with utmost confidentiality, adhering to relevant data protection laws and guidelines. We prioritize

transparency in data handling practices to ensure users' trust and confidence in the app's privacy safeguards.

Furthermore, inclusivity and accessibility are key considerations in our app's design. We prioritize accessibility features, ensuring individuals of diverse abilities can navigate and engage with the application seamlessly. Additionally, we aim to address language and cultural considerations, making the app accessible to a global audience. Our commitment to inclusivity extends to ensuring that the app's features cater to users from various backgrounds and communities.

The app is not just a technological tool; it serves as a platform for education and awareness. Our ethical responsibility transcends technology and includes promoting empathy and responsible pet ownership within communities. We are committed to providing accurate and unbiased information about the challenges faced by stray dogs, aiming to raise awareness and foster a sense of responsibility towards animal welfare.

Moreover, collaboration with local animal welfare organizations is central to our ethical approach. We strive to work in tandem with existing efforts, respecting established protocols, and avoiding duplication of services. By collaborating closely with shelters, rescue organizations, and authorities, we ensure that our technology complements and enhances overall efforts for stray animal welfare.

Stray Dog Tracking App's ethical framework guides its development and implementation, ensuring the app serves as a responsible, effective, and humane tool in the endeavor to improve the lives of stray dogs and the communities they inhabit.

4 FUTURE WORK, CONCLUSION AND LIMITATIONS

4.1 Future work:

Addressing these limitations warrants further exploration and strategic planning. Future efforts could focus on enhancing GPS technology robustness, promoting community engagement strategies, implementing measures to improve the accuracy of reported sightings, and fostering collaboration between app users and animal shelters. Furthermore, devising solutions to mitigate privacy concerns and exploring cultural attitudes toward pet ownership could enhance the app's functionality and adoption.

4.2 Conclusion:

The Stray Dog Tracking App exhibits notable limitations impacting its efficiency in managing stray animals. These challenges underscore the need for comprehensive strategies addressing technological, user engagement, resource allocation, ethical, and cultural factors. Overcoming these limitations through targeted improvements and strategic collaborations is crucial to maximize the app's potential in effectively addressing the complexities of stray animal management.

4.3 Limitations:

The accuracy of reported sightings emerges as a critical concern, with potential misidentifications or mistaken reports diverting resources away from genuine stray dog issues. Additionally, collaborating with local animal shelters faces hurdles due to resource constraints, including limited staffing, financial resources, and differing organizational priorities. Ethical considerations related to privacy implications of GPS tracking for stray dogs further complicate the app's functionality.

These limitations highlight the complexities involved in managing stray animals effectively. Challenges in user engagement, resource availability, and ethical considerations underscore the multifaceted nature of addressing stray animal concerns. Misidentifications and resource constraints may hamper the app's intended efficacy, necessitating careful consideration and strategic approaches to overcome these hurdles.

REFERENCES

[1] Stray Animal Population Control: Methods, Public Health Concern, Ethics, and Animal Welfare Issues - Scienceline Publication Repository. (n.d.). http://eprints.science-line.com/id/eprint/352/

[2] Mohanty CR, Radhakrishnan RV, Jain M, Sasmal PK, Hansda U, Vuppala SK, Doki SK. A Study of the Pattern of Injuries Sustained from Road Traffic Accidents Caused by Impact with Stray Animals. J Emerg Trauma Shock. 2021 Jan-Mar;14 (1):23-27. doi: 10.4103/JETS.JETS_29_20. Epub 2021 Mar 23. PMID: 33911432; PMCID: PMC8054802.

[3] Gareth Davey, Xiang Zhao & Mei Mei Khor (2020) Heterogeneity in beliefs about feeding stray animals: the complexity of human–animal interaction, Human Dimensions of Wildlife, 25:1, 100-103, DOI: 10.1080/10871209.2019.1692099

[4] Anisimov, Aleksey Pavlovich and Ryzhenkov, Anatoliy Jakovlevich. "Is it possible to change the destiny of stray animals by legal means?" International Journal of Legal Discourse, vol. 4, no. 2, 2019, pp. 143-166. https://doi.org/10.1515/ijld-2019-2020

Artificial Intelligence based Face Recognition and its Applications in Effective Governance

Dr. Xavier Chelladurai

ABSTRACT

The influence of Artificial Intelligence (AI) has instigated a profound shift in human existence, fundamentally altering the way we operate. This transformation encompasses the augmentation of decision-making processes through cutting-edge technologies like Machine Learning, Deep Learning, and other domains where machines adeptly perform tasks that were once solely human-centric. Notably, Face Recognition has emerged as a pivotal domain significantly impacted by AI advancements. This study proposes an innovative approach to Face Recognition leveraging Deep Learning models. The core methodology involves subjecting each image to a comprehensive feature extraction process using Deep Learning techniques, resulting in the creation of a distinctive 128-number vector known as the "face signature." This signature encapsulates diverse facial features essential for recognition. The crux of face comparison relies on evaluating the resemblance between two faces, a determination accomplished through measuring the Euclidean distance between their respective face signatures. This distance metric serves as the cornerstone for establishing facial similarity and dissimilarity. To substantiate the proposed approach, rigorous testing has been conducted utilizing the Python Face Recognition library. The empirical validation aimed to assess the effectiveness and reliability of the developed methodology in accurately identifying and distinguishing between faces based on their unique signatures.

Dr. Xavier Chelladurai
Professor, Department of CSE, School of Engineering and Technology,
CHRIST (Deemed to be University), Bangalore, India.
Email: xavier.c@christuniversity.in

1. INTRODUCTION

Artificial Intelligence is transforming every walk of human life. Researchers who study human civilization believe that the AI transformation is going to be more rapid than all other transformation we have had so far. Soon, we expect robots with the human capabilities coexist with us. The robots can see, hear, feel, recognize, analyze, and also work creatively. Face recognition is a branch of Artificial Intelligence technology. Face recognition is about detecting a human face in an image, recognize the identity, gender, age and more attributes of the person and even the emotions. Face recognition systems are developed using various approaches.

Face recognition is one of the high-level intellectual capabilities of human beings. Once when I was walking in the crowded cafeteria of my office, a girl met me and greeted me “Hello! Uncle”. In a fraction of a second, I was able to recognize that it is my daughter's classmate. I had last met her 25 years ago as a 10-year-old girl. Now after 25 years, I could recognize her face in a fraction of a second. We do not know how exactly our brain recognizes, but we are able to recognize. We are going to study how this capability can be built in a machine.

Let us first study about face detection. Face detection is about observing an image and detecting the human faces in it. A human face has a specific structure with the nose, chin, cheek, forehead, hair, moustache, beard, eyes, eyebrow, eyelashes, mouth, lips, and several other components and features. As we know a human face and its components, we can detect a human face in an image. In a group photo taken at an outdoor with animals, flowers, birds, and other objects, we detect all the faces in it. The problem of viewing an image and detecting all human faces in the picture and extracting them in the form of rectangular frames is called face detection.

Face Recognition is a higher level of intelligent activity than face detection. Face recognition is about reading a human face either directly or in an image and recognize one or more of the following:

- o Identity of the person
- o Age
- o Gender

- o Ethnicity
- o Attractiveness,
- o State of mind,
- o Emotions, anger, fear, happy,
- o Trustworthiness
- o Competence
- o Dominance
- o Extraversion,
- o Mental Health

Enormous amount of research is happening by leading Artificial Intelligence companies such as Google, Amazon, Facebook, IBM etc. Hundreds and thousands of startup companies in the bay area of US, Continental Europe, Israel, Bangalore and many more places are working on great applications in this area.

2. BIOMETRICS & FACE RECOGNITION

Right form early days, scientists are trying to uniquely identify human being based on one or more parts of the body and their measurement. Biometrics is the study of body measurements and computation of ratios and other relationships among them. This is used to see if we can uniquely identify people from these measurements. Fingerprints and IRIS have been successfully studied and used for several years in medical research, immigration, travel documentation, criminology, educational certification and many more.

Medical researchers have been studying how human brain stores faces and recognize them. They have found that the brain does the following cognitive activities:

- o Face Detection
- o Face categorization based on gender, ethnicity, age etc.
- o Face discrimination
- o Faced Individuation
- o Face memory
- o Face naming

Medical Researchers are working on a few interesting concepts such as Face Diet and Other Race Effect (ORE) areas in Face Recognition. Face recognition capability varies widely among the people. Some people are super recognizers. Even if they meet somebody once, they recognize the face even after a long time. Some people are very poor in face recognition. They require a great deal of training and practice to familiarize a face. There are people who have a very high level of inability to recognize faces. This inability is medically diagnosed as Face blindness or Developmental prosopagnosics (DP). In order to improve face recognition capabilities, the researchers propose a few activities.

Face Diet is the number of faces we come across every day. People who have a rich face diet are found to have a high face recognition capability. People who live meet very few people in a day. So, they have poor face diets. They have a poor face recognition capability.

Other Race Effect (ORE): We come across people from different ethnicity and race. There are some common face features for every race. When people have no familiarity with a particular race, they find it hard to distinguish faces of that race. All the faces of that race look alike for them. As we become familiar with more and more people of that race, the ability to distinguish increases. This is called Other Race Effect (ORE). When I watch movies from Japan, China or Korea, I have experienced difficulty in recognizing different characters. For me all the

Medical Researchers also work on Nose recognition, Ear Recognition, Mouth Recognition etc. Critical Features of face are studied in detail.

3. FACE RECOGNITION APPROACHES

Broadly there are three sets of scientists working on face recognition and related fields. They are

- o Medical Scientists – Cognitive strategies
- o Anthropological Researchers
- o Artificial Intelligence Researchers using Computer Vision with Deep Learning Convolution Neutral Networks.

In Medical Research, Scientists work on two broad approaches.

1. Holistic Face Processing in the brain. In the approach scientist work on how the brain stores and retrieves data.
2. Recognition of parts such as nose, mouth etc. The work on various shapes of the parts and specialize on recognizing them. You can observe that the noses have various shapes. Mouth and lips are of different shapes.

In Biological Anthropology, researchers study the physical metrics of people from various ethnicities. Craniofacial Anthropometry is the study of head and face measurements.

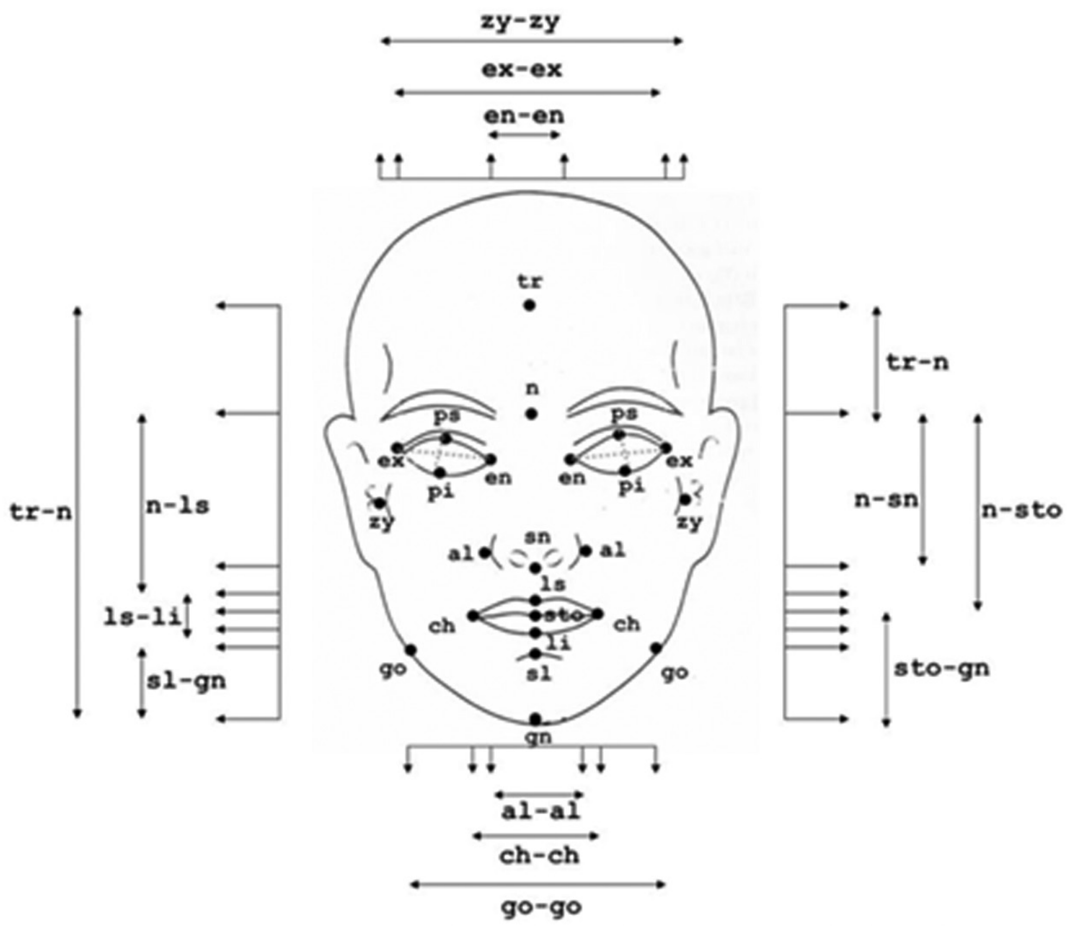


Figure 1 Metrics studied Craniofacial Anthropometry

4. ARTIFICIAL INTELLIGENCE RESEARCH

In Artificial Intelligence research the image of the face is represented as a matrix of pixels. By analyzing the image, we try to find the unique set of numbers, called a signature. We use face signature for face recognition. Each image is considered as a matrix of pixels. Each pixel is represented by a set of numbers.

If the image is BW, the content of a pixel is represented as a number from 0 to 255. Colors are represented by the combination of Red - Green - Blue combination. So, each pixel is represented a tuple (r, g, b) where each entry is a number between 0 and 255.

5. FACE SIGNATURE

Face Recognition Algorithm works very similar to any common Solution Searching algorithm. Given an array of numbers a_0, a_1, a_2, a_3,… a_{n-1}, and a token X, search for X in the array. In other words, find the index i where, $X = a_i$. Whenever we succeed in a specific ^{i}, we output that X is found in index i. In a Face Recognition based Employee attendance management system, the face photo of all the employees is taken during their joining process and stored as p_0, p_1, p_2, p_3, …. p_{n-1}. Here, the index i represents the employee number. The names are also stored in another array name.

Index	0	1	2	…	n-1
name	Ram	Sundar	Seetha		Kannan
Photo					

In the morning when employees arrive in the office/ factory, they stand in front of a camera. The camera captures the photo as X. Now X is compared with photos p_0, p_1, p_2, p_3, ….p_{n-1}. If $p^i = X$ for a specific i, the attendance entry is made with time stamp for the employee ID i.

Now let us see how $P_i = X$ is computed as TRUE or FALSE. P_i and X are two image files. Each image file is a matrix of pixels. Each pixel has one number if it is a BW photo and three if it a color photo. Comparing p_i and X pixel by pixel is complex as well as meaningless.

The photo pi of the employee was captured during the joining process. But X is captured on the day of attendance. So, X and pi are not same, pixel by pixel. To solve this problem, lot of research has happened in the past 50 years or so.

Consider the recognition of flags. When the flags of all the countries are given (Figure 3), how do we recognize the flag of a specific country? We know that the flag of Canada in as given below in Figure 2:



Figure 2 Flag of Canada

When you are shown a picture X and asked to check if it is a Canadian flag, are we comparing X and the Canadian flag pixel by pixel? No when we observe image X, we just observe some important features such as colors, shapes insides the flag, etc. This is called feature extraction.

Based upon these important features, we create a vector of X as [x0, x1, x2, … xd-1]. This is called the signature vector of X, and d is called the dimension. For flag recognition problem, the dimension may be a small number, may be d= 8 or d = 4. This means that by observing just 8 features, we can recognize a flag.



Figure 3 Flags of some countries

A human face is much more complex than a national flag. To recognize the human face, we may need to extract 128 features or 256 features. We get the feature of pi as [pi0, pi1, pi2, …. pid-1]. This is called the face signature of face pi.

When we need to compare two faces, we compare their signatures. If the two face signatures are approximately same, we conclude that the two images represent the face of the same person.

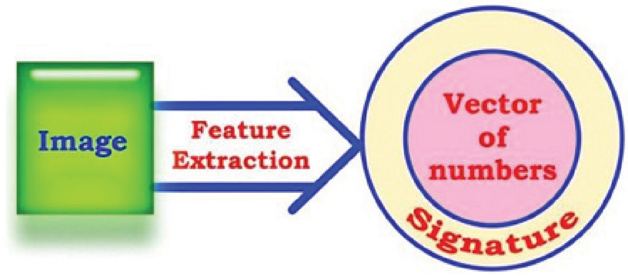


Figure 4 Signature Vector

Convolution Neural Network

The image of human face has a standard format, with one nose, two eyes, two ears, a chin, cheek etc. So, how do we want to distinguish human faces? We must do feature extraction and create a Face signature. In Deep Learning, Convolution Neural Networks are designed and trained using data set having thousands of human faces. This CNN outputs the face signature as a vector of dimension 128 or 256. The face signature is a vector of numbers. To compare if two faces represent the same person, we compare the two face signatures and check if they are almost same.

Example

Consider the three photos as shown in Figure 2.3. Photo P0 is a 300 x 400-pixel color image. This is represented by a matrix of 300 x 400 pixels. Each of these 1,20,000 pixels is represented by a tuple (r, g, b) where each of r, g, b, are numbers between 0 and 255. So, we represent the phot P0 as 3 x 1,20,000 = 3,60,000 numbers. We first convert the photo into a vector (Face Signature) Similarly, we have photo P1 and P2. These images can be converted into their respective signatures FS0, FS1, and FS2 with d=128 or 256 each.

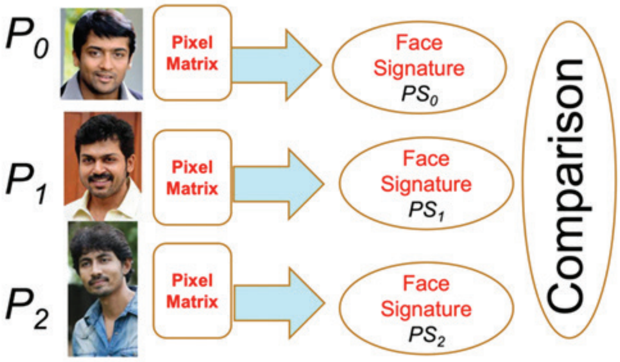


Figure 5 Photos of Faces and Face Signatures

Consider the photo of a known person Pk. The face signature is already computed and stored as PSk (k represents known people). Now if we are given an unknown image Puk (uk means Unknown), we first compute the signature of the unknown image as PSuk. We compare PSk and PSuk. If they are almost (approximately) equal, we recognize that the image PUK is that of Pk (Known already). This is illustrated in Figure 6.

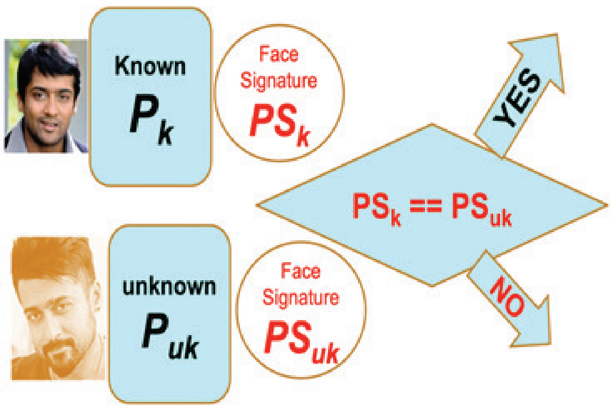


Figure 6 Compare Signatures and recognize faces

6 APPLICATIONS IN EFFICIENT GOVERNANCE

Artificial intelligence (AI) has emerged as a powerful tool in various sectors, and its application in face recognition technology has garnered significant attention, particularly in governance. The utilization of AI-based face recognition for efficient governance purposes has immense potential to revolutionize various aspects of public administration, law enforcement, and service delivery.

Governments worldwide are exploring the integration of AI-powered face recognition systems into their governance frameworks. One of the primary applications is in law enforcement and public safety. These systems enable authorities to identify individuals swiftly and accurately, aiding in the apprehension of suspects, locating missing persons, and enhancing overall security. By leveraging AI algorithms capable of analyzing facial features, these systems can match faces against existing databases, providing law enforcement with valuable tools for investigations.

Moreover, AI-powered face recognition can streamline administrative processes. In areas like identity verification for government services, social welfare programs, or voting systems, implementing facial recognition technology can enhance accuracy, reduce fraud, and ensure that services are efficiently allocated to eligible individuals. For instance, in government offices or border control, facial recognition can expedite identity verification, reducing wait times and enhancing the overall efficiency of administrative procedures.

Efficient governance also involves ensuring public safety in various spaces. AI-driven face recognition technology can be integrated into security systems in public areas such as airports, train stations, and stadiums to enhance surveillance capabilities. It enables real-time monitoring and identification of individuals on watch lists or those posing potential security threats, contributing to preemptive measures and crisis management.

However, while AI-based face recognition offers several benefits, its implementation raises ethical and privacy concerns. One significant challenge is the potential misuse of this technology, leading to mass surveillance or infringing upon individuals' privacy rights. To address these concerns, governments must establish robust regulatory frameworks and guidelines governing the use of facial recognition technology. Clear policies should delineate permissible uses, data protection measures, limitations on data retention, and mechanisms for obtaining consent and ensuring transparency.

Moreover, the accuracy and biases inherent in AI algorithms used for face recognition pose another challenge. These systems might exhibit biases concerning race, gender, or age, leading to erroneous identifications and potential discrimination. Continuous refinement of these algorithms and rigorous testing against diverse datasets are crucial to mitigate biases and ensure fair and accurate outcomes.

Another consideration is the need for public awareness and education about the capabilities and limitations of AI-powered face recognition. Building trust among citizens regarding the responsible use of this technology is vital for its acceptance and successful integration into governance frameworks.

7. CONCLUSION

AI-based face recognition holds immense promise for efficient governance across various domains, from law enforcement and administrative processes to public safety and service delivery. However, its implementation requires a careful balance between leveraging its potential benefits and addressing the ethical, privacy, and accuracy concerns associated with its deployment. Through thoughtful regulation, ongoing technological advancements, and public engagement, governments can harness the power of AI-based face recognition to enhance governance while safeguarding individual rights and privacy.

REFERENCES

[1] N. Ramanathan, R. Chellappa, Modeling Age Progression in Young Faces, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2006) (CVPR'06)

[2] Vahid Kazemi and Josephine Sullivan, One Millisecond Face Alignment with an Ensemble of Regression Trees, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1867-1874

[3] Florian Schroff, Dmitry Kalenichenko, James Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering, CVPR 2015 Google Publication

[4] C. G. Zeinstra1, D. Meuwly, A. C. C. Ruifrok, R. N. J. Veldhuis1, L. J., Spreeuwers, Forensic Face Recognition as a Means to Determine Strength of Evidence: A Survey, Forensic Sci Rev 30:21–32; 2018, pp 22-31

[5] Narges Shafieian, Ali Amiri, Recognizing Facial Expressions by Using New Algorithm Based on Combined Approach of PSO and MLBP, American Journal of Computer Science and Engineering, 2019; 6(1): 10-15.

[6] Ipek Oruca,b,!, Benjamin Balasc, Michael S. Landyd, Face perception: A brief journey through recent discoveries and current directions, Vision Research 157 (2019) pp 1-9.

[7] Zipporah Jiya, James Gana Josiah, Anthropometric study of craniofacial morphology among Nupe ethnic group in Niger State, Nigeria, Elsevier Forensic Science International: Reports, Volume 6, December 2022, 100291

Assistive Tool for Converting Sign Language to Tamil Text and Speech for Individuals with Hearing and Speech Impairment using Deep Learning

Dr. G. Indirani, Dr. G. Revathy, Dr. B. Kumaravel

ABSTRACT

Hearing loss affects more than 5% of the population, according to the WHO. Many of these people have never been exposed to sign languages, and it has been noticed that learning to sign to connect with others by expressing their affection or feelings provides significant psychological comfort. Deaf and dumb persons primarily utilize sign language to communicate their thoughts and ideas to others around them through various hand and body motions. As a result, we created an assistive program that recognizes hand gesture components and converts them to Tamil voice and text that a normal person can readily comprehend, and similarly, normal people's movements are intelligible by Deaf and Dumb people. The assistive device enhances the self-esteem of physically challenged people.

Dr. G. Indirani,
Associate Professor, Department of CSE, Government College of Engineering, Thanjavur.
Dr. G. Revathy,
Assistant Professor, Department of CSE, SRC, SASTRA Deemed University, Kumbakonam.
Dr. B. Kumaravel,
Associate Professor, Department of Civil Engineering, Annamalai University, Chidambaram.

1. INTRODUCTION

The exact moment that sign language was introduced is not known, although there were different ages at which different people in different places connected with one another. In addition to more than 6000 spoken languages that make up our global community, there are numerous sign languages that are used with one or both hands. American Sign Language (ASL) is an example of a single-hand user, while Indian Sign Language (ISL) is an example of a double-hand user. Over 300 sign languages are used throughout the world, according to the UN. The Indian Sign Language (ISL) is an additional example. Sign language is a language that is expressed using hand movements, facial emotions, body postures, and other gestures. Its principal goal is to improve people's communication skills, especially for people who have hearing impairments or speech issues. Sign language can offer a more comprehensive degree of context and context understanding being conveyed due to the integration of facial hand movements and facial expressions in the conversation method. In India, the quantity of people who have hearing impairments is estimated as 18 million people, and forty one in every thousand children is affected by severe loss of hearing.

Speech recognition is the process by which an automated system identifies the speech of a speaker. These systems may rely on a speaker or not at all. Speaker-independent systems train on spoken utterances of the relevant isolated words made by a single set of speakers, while testing makes use of phrases produced by a different collection of speakers. Speaker-dependent systems would employ a different set of utterances for testing and a single set of utterances made by all speakers for training. Speech recognition systems that are not dependent on the speaker uttering the words or sentences can be used to identify and comprehend them.

Speech recognition systems that rely on the speaker find use in controlling speaker-specific functions. ISL uses both manual and non-manual communication techniques to convey concepts, sentiments, and body language. ISL symbols can be broadly categorized into three types: one-handed, two-handed, and non-manual signs. We call a sign made by a signer with their hands to communicate with others a “manual sign.” Regardless of whether the sign is made with one or two hands,

this is the case. Non-manual signals provide meaning through changes in posture, facial expression, and emotional state without the use of hands. It's evident that some alphabets—like C, I, J, L, O, U, V, and W—can be represented with just one hand, while the remaining alphabets need the use of both hands.

Caretakers may find it extremely challenging to comprehend the distorted and disordered speech of individuals affected by hearing impairment, dysarthria, autism spectrum disorder, or speech impairment in order to provide the necessary support. Individuals diagnosed with hearing impairments range in severity from modest to significant hearing loss. Youngsters with modest hearing loss would not be able to interpret what other people are saying since they would not comprehend how speech is produced. Youngsters with minor hearing loss are unable to distinguish between various unvoiced sounds, such as "the," "f," "s," "t," and "sh." To ensure that communication is understood clearly, certain sounds need to be prioritized and spoken. Even for average persons, it is difficult to recognize speech if the high-frequency unvoiced sound is absent. The images in Indian Sign Language are given in Figure 1.

From the perspective of study, regular dialogue or communication between the HI and regular people may not be beneficial and continues to be difficult. These kids can pick up speaking and conversing with people quickly if they receive regular and consistent training. They are socially adept and do not need help from others.

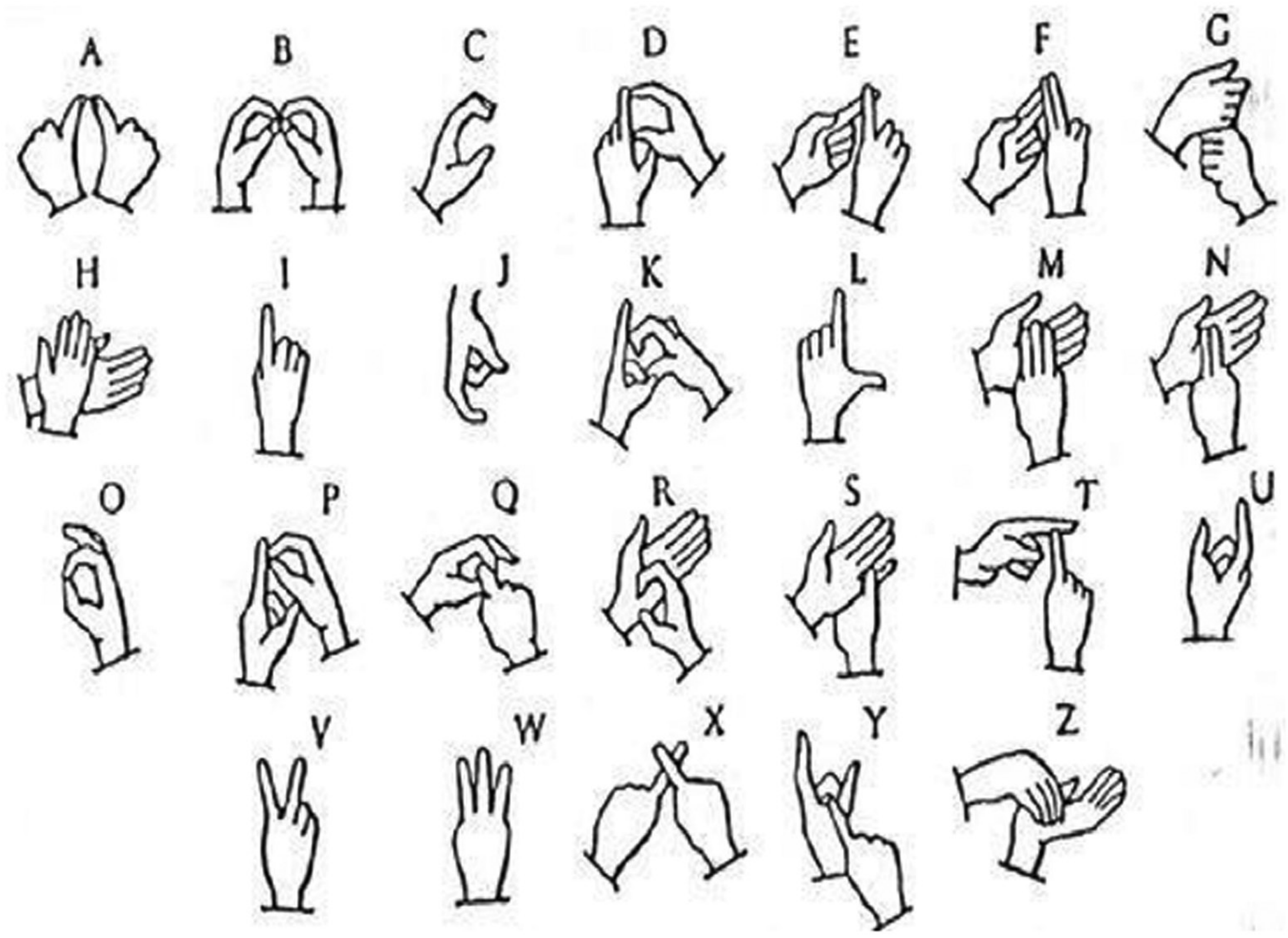


Figure 1: Alphabets in Indian Sign Language for English Alphabets.

With consistent instruction, it is possible to help HI interpret and perceive speech because their vocal tract structure is similar to that of typical people. Adults experience hearing impairment in addition to arthritis and hypertension.

The work is been categorized as follows the topic 2 will be literature survey, 3 is dataset description, 4 is methodology been used, 5 results and discussion followed by conclusion.

2. LITERATURE SURVEY

Author says that the computer system is creating a model of speech synthesis that incorporates a number of natural language processing features. Articulatory, formant, and concatenate synthesis are the methods used in speech synthesis exploration. These methods result in an exponential increase in the error rate throughout the system's operation and greater aperiodic distortion. In an effort to obtain higher performance, recent

advancements in voice synthesis have made significant strides toward deep learning. Large-scale data is leveraged to provide voice synthesis with effective feature representations. This study article's major goal is to apply deep learning techniques to speech synthesis and evaluate the results in terms of aperiodic distortion by comparing the results with earlier models of natural language processing algorithms [1].

Authors discusses that Text-to-Speech (TTS) research has made significant strides in creating lifelike speech with the introduction of deep learning. Modern TTS systems provide average prosody, which lacks the range and expressiveness inherent in human speech. The other model employs a variational autoencoder, which is supplemented by normalizing flows and an adversarial training procedure. We trained three internal Bangla (Bengali) datasets with variable quantities of expressive talks. We present a comparative analysis on the influence of expressive voice percentage in training data. Both subjective and objective assessments show that the suggested models outperform the baseline autoregressive Tacotron2 [2].

Authors examine a Deep Neural Network-based Text-to-Speech synthesis for Arabic. To evaluate the system, subjective and objective tests were performed. For subjective evaluation, we employed the Mean Opinion Score (MOS), and the Diagnostic Rhyme Test (DRT) to assess the intelligibility of particular consonants and vowels. [3].

Authors discuss that Accent recognition systems have grown in relevance as a result of globalization and the emergence of voice-activated devices. Foreign-accented English has distinct acoustic properties than native English. It differs depending on the native language of the speakers [4].

Authors suggests that the growing number of social media users, as well as the manner in which these users communicate in regional languages, has prompted experts to consider the uniqueness of these languages. Native speakers are concerned that the regional languages may lose their uniqueness over time. As a result, checking for accuracy when writing is an essential problem in the field of natural language processing. The traditional techniques for determining the soundness of a phrase include the application of numerous syntactic and semantic criteria, which are limitless due to the Assamese language's unrestricted word order [5,8,9].

The Authors claims that the goal of their study is to design and build an accurate voice recognition system for a collection of predetermined words derived from short audio samples. It employs Google's TensorFlow's The Speech Commands Dataset v0.01. Voice user interfaces with key-word spotting can leverage isolated word speech recognition. The ultimate aim is to identify and recognize 10 words, as well as create classes for “unknown” words that are distinct from the “silence” class. Acoustic noise and differences in recording conditions are two of the issues that current speech recognition technology faces. MFCCs and Mel-spectrograms were employed to extract relevant information from the signal. Convolutional neural network (CNN) was employed for classification [6,7,10].

Authors says that brand-new wearable gadget with a camera for voice augmentation and tracking across a 360° range in conjunction with an array of microphones was suggested by the research. To monitor the speech target in a flexible way, the suggested sensing approach merged voice and computer vision (CV) methods. Students with hearing loss can focus more readily during class instruction when the device can scan a specific voice source over 360° and eliminate interference-related noise, improving the auto scan hearing-impaired procedure[11,12].

Authors suggest that CCi-MOBILE is a computationally robust signal processing testing platform designed for researchers working with the hearing-impaired community. This paper describes its conception, development, clinical assessment, and applications. Researchers may create and evaluate complicated speech processing algorithms offline and in real time using the specially designed portable research platform. Compatibility with Cochlear Corporation implants, it may be operated with user-friendly, open-source software. Results of an acute trial with implant users' speech intelligibility in quiet and loud environments are presented, and then the FPGA design and hardware processing pipeline for CI stimulation is explored. When compared to the clinical processors of CI users, the findings consistently demonstrate a level of performance that validates the platform's feasibility in investigations including chronic CI [13].

Authors suggests considering articulatory feature (AF) sequences of phonemes as multi-label objects in speech spectrograms in order to identify AFs from spoken utterances using object identification algorithms. Localizing a series of multi-label AFs in a speech signal is what the suggested system, named AFD-Obj, does. The two main stages of AFD-Obj are as follows: first, we formulate the AFs detection problem as an object detection problem and prepare the data to satisfy the requirements of object detectors by producing a spectral three-channel image from the speech signal and an annotation for each utterance. Secondly, the suggested system is trained to identify AF sequences and their boundaries using annotated pictures[14].

Authors examines the influence of the system's correctness on the number of phoneme states, learning rate, and training set value—three key aspects of voice

command pronunciation models. It is suggested to use a neural network-based voice recognition system. It needs knowledge of the fundamentals of voice recognition to operate a speech recognition system, which is not a simple task. Google Speech Recognition and Pocket Sphinx are compared with the created system. With an accuracy rate of 84.4%, the suggested system can identify voice instructions [15].

3. METHODOLOGY

There are 5 stages in the proposed methodology. They are

Stage 1 Collection of images (Hand Gestures)

Stage 2 Image Pre-Processing and Edge Detection

Stage 3 Feature Extraction

Stage 4 Classification using Deep learning models

Stage 5 Conversion of Gesture to Tamil text and Tamil Speech.

The architectural diagram of the proposed methodology is given in Figure 2. In stage 1, the images are obtained from the end user for the evaluation purpose. In stage 2 the basic pre processing of images are done by pixel to pixel then an edge detection algorithm is been applied (Prewitt). Stage 3 is feature extractions of images followed by stage 4 Deep models (Inceptionnetv2) are applied in identification of data. Then last stage is conversion of output to Tamil language and Tamil speech.

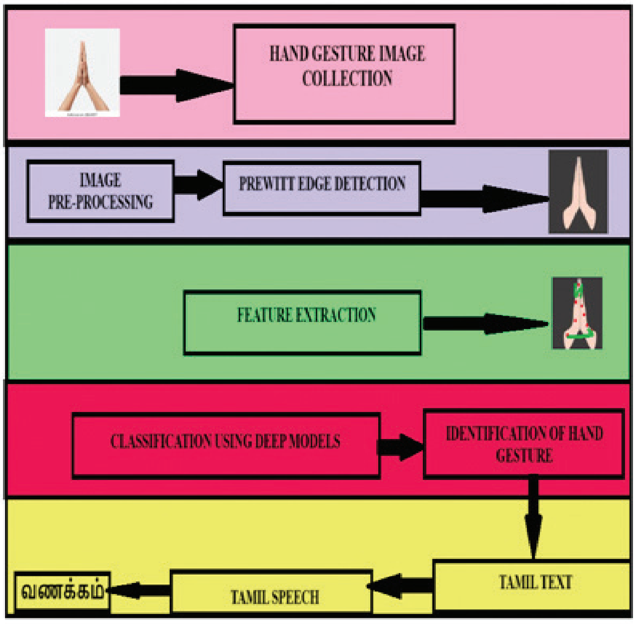


Figure 2: Architecture Diagram of the Proposed Model

The Deep model we use here for identification of Hand Gesture is InceptionnetV2. The architecture is described in Figure 3.

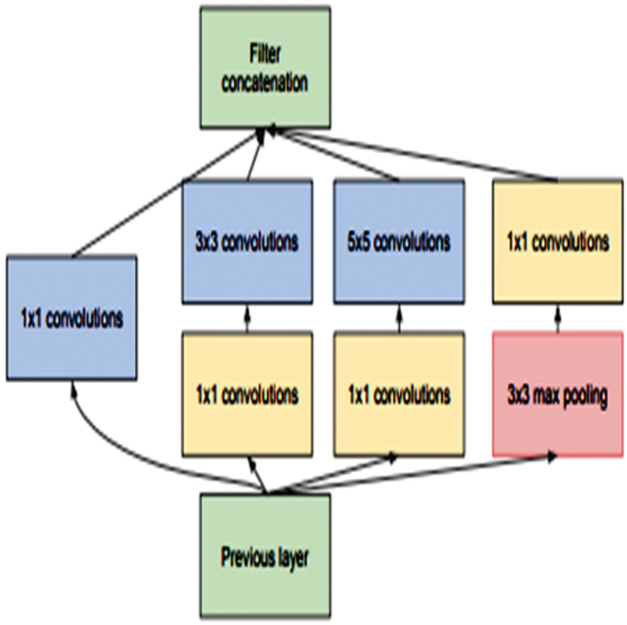


Figure 3: Inception netV2 architecture

Figure 3 explains the deep model provided with a Filter for the concatenation and there are 6 various convolution layers and a max pooling layer which makes the result more accurate and effective.

Once the results are been obtained from the deep model then it is converted to speech using GTTS.

4. RESULTS AND DISCUSSIONS

There are more than 3500 images have been trained with various kids in and around Kumbakonam regime. Out of which some images are given for testing and the results obtained are shown

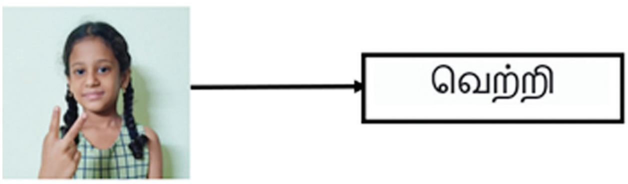


Figure 4 Identification of Single Hand gesture

Figure 4: explains the single hand gesture testing for the word and the exact word is been predicted.

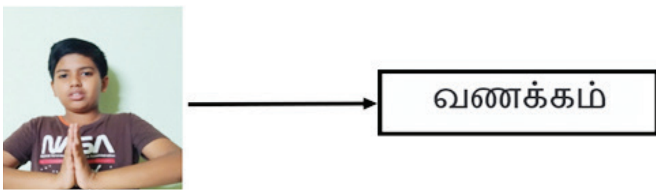


Figure 5: Identification of Double hand gesture

Figure 5 explains the double hand gesture testing for the word and the exact word is been predicted.

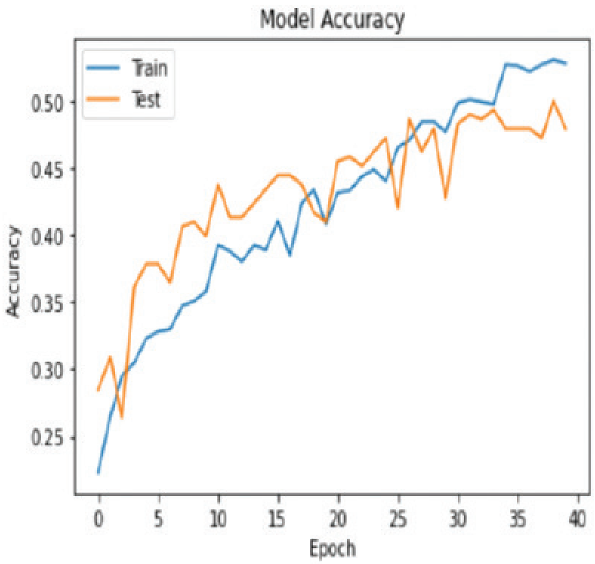


Figure 6: Training model accuracy for InceptionnetV2.

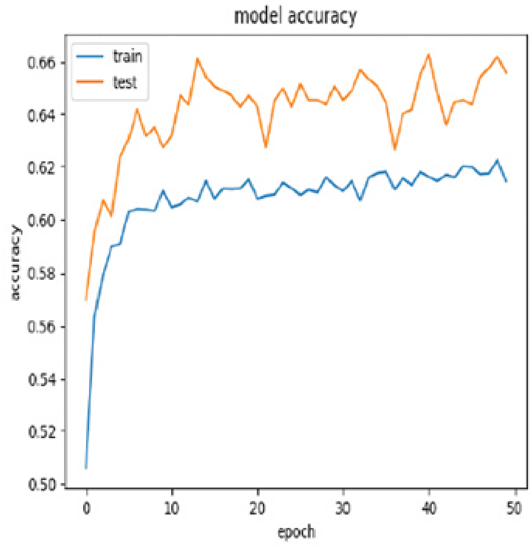


Figure 7: Testing model accuracy for InceptionnetV2.

Figure 6 explains the accuracy of the model for various epochs and Figure 7 is the testing result for various epochs.

CONCLUSION

The assistive tool is been developed which would assist the hearing-impaired persons and Dumb persons for an effective communication like normal persons. Although the testing results are good when compared with state of art models yet it has to be improved a lot for many other hand gestures and the communication to Tamil and also needs an well effective development. The assistive tool is more planned for effective training with much many more complicated images and converting it as an web application or mobile application in upcoming days.

REFERENCES

1. Adam, Dr Edriss Eisa Babikir. "Deep learning based NLP techniques in text to speech synthesis for communication recognition." Journal of Soft Computing Paradigm 2.4 (2020): 209-215.
2. Ahmad, Arif, et al. "Expressive Speech synthesis by modeling prosody with variational autoencoders for bangla text-to-speech." (2022).
3. Amrouche, Aissa, et al. "DNN-Based Arabic Speech Synthesis." 2022 9th International Conference on Electrical and Electronics Engineering (ICEEE). IEEE, 2022.
4. Barkana, Buket D., and Aarchi Patel. "Analysis of vowel production in Mandarin/Hindi/American-accented English for accent recognition systems." Applied Acoustics 162 (2020): 107203.
5. Bhuyan, M. P., S. K. Sarma, and M. Rahman. "Natural language processing based stochastic model for the correctness of assamese sentences." 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2020.
6. Chen, Li-Wei, and Alexander Rudnicky. "Fine-grained style control in transformer-based text-to-speech synthesis." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.
7. Du, Chenpeng, et al. "VQTTS: High-fidelity text-to-speech synthesis with self-supervised VQ acoustic feature." arXiv preprint arXiv:2204.00768 (2022).
8. Fahmy, Fady K., Hazem M. Abbas, and Mahmoud I. Khalil. "Boosting subjective quality of Arabic text-to-speech (TTS) using end-to-end deep architecture." International Journal of Speech Technology 25.1 (2022): 79-88.
9. Joshi, Sonal, et al. "Study of pre-processing defenses against adversarial attacks on state-of-the-art speaker recognition systems." IEEE Transactions on Information Forensics and Security 16 (2021): 4811-4826.
10. Soliman A, Mohamed S, Abdelrahman IA (2021) Isolated word speech recognition using convolutional neural network, 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering(ICCCEEE), pp. 1–6, https:// doi. org/ 10. 1109/ ICCCE EE496 95. 2021. 94296 84.
11. Bhat, Gautam S., et al. "A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone." IEEE Access 7 (2019): 78421-78433.
12. Du, Yi-Chun, et al. "A Wearable Device of 360° Speech Tracking and Enhancement for Hearing Loss Students in Class." IEEE Sensors Journal 22.21 (2022): 21163-21171.
13. Ghosh, Ria, Hussnain Ali, and John HL Hansen. "CCi-MOBILE: A portable real time speech processing platform for cochlear implant and hearing research." IEEE Transactions on Biomedical Engineering 69.3 (2021): 1251-1263.
14. Algabri, Mohammed, et al. "Deep learning-based detection of articulatory features in arabic and english speech." Sensors 21.4 (2021): 1205.
15. Lyashenko, Vyacheslav, et al. "Recognition of voice commands based on neural network." (2021).

DISTINCT: Deep Identification of Tamil Language Speech through Modified Features and Neural Networks

Kanimozhi Suguna S, Prema S, Vasanthakumari M

ABSTRACT

This study will present a model based on Deep Neural Networks to identify various regional dialects within the Tamil language. In this context, an idiom refers to a specific variation of the language unique to a particular district or social community. While many researchers traditionally employ Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) for speech-related or processing applications, the contemporary landscape favors the integration of neural networks across various domains, yielding robust results. Understanding the acceptable decisions within the architecture of Deep Neural Networks (DNN) is crucial for enhancing the performance of state-of-the-art speech recognition systems. Our research delves into determining which aspects of the DNN audio model are most impactful on the performance of a speech recognition system with the implementation of a Convolution Neural Network (CNN) hybrid with Long-Short Term Memory (LSTM) termed CNN-LSTM. Concentrating on a feed-forward system effectively identifies dialects when combined with modified Mel-Frequency Cepstral Coefficients (MFCC) features by creating a model CNN-LSTM-MFCC. Also, the DNN model will be applied to identify three regional dialects within the Tamil language—namely, those of Northern Tamil Nadu, Southern Tamil Nadu, and Eastern Tamil Nadu—using MFCC features. Upon comparison with traditional HMM and GMM models, our findings indicate that the DNN (CNN-LSTM-MFCC) model yields superior accuracy in identifying these dialects.

Kanimozhi Suguna S, Assistant Professor, Department of Computer Applications
Prema S, Assistant Professor & Head, Department of Computer Applications
Vasanthakumari M, Assistant Professor & Head, Department of Tamil
Arulmigu Arthanareeswarar Arts and Science College, Tiruchengode, Namakkal.

1. INTRODUCTION

1.1 Role of Speech Identification in Tamil Language

Identifying speech in the Tamil language is highly significant for multiple reasons. Tamil is a classical language with a robust literary past. It includes various dialects that represent its historical and cultural diversity. Precise speech recognition aids in conserving and comprehending these dialectical subtleties, supporting the broader endeavors of language and cultural preservation. Furthermore, speech recognition is crucial in advancing Tamil's resilient natural language processing systems, encompassing automated transcription services and voice-activated technologies. This technology not only improves accessibility but also creates opportunities for creating region-specific applications that cater to the different linguistic landscapes of Tamil speakers.

Moreover, identifying speech in Tamil is essential for linguistic research, as it aids linguists and researchers in deciphering the complex linguistic patterns and variations that define the language. This research eventually enhances our comprehension of Tamil linguistics. To summarize, speech recognition in Tamil is not just a technological achievement but also a means to preserve cultural variety, improve communication, and promote language understanding within the Tamil-speaking community.

1.2 Objective of the Study and Modified features of Deep Neural Network (DNN)

This study aims to enhance Automatic Speech Recognition (ASR) in the Tamil language for vulnerable populations, including elderly adults and transgender individuals, by utilizing Deep Neural Networks (DNNs) with updated characteristics. The main goal is to address the specific challenges faced by these groups while acknowledging potential variations in speech patterns due to factors such as age-related [1] changes or transgender identity. The choice to employ Deep Neural Networks (DNNs) signifies a contemporary approach to leveraging advanced machine learning methodologies, well-known for their ability to acquire complex patterns and representations from data [2]. The authors emphasize the necessity of modifying

attributes, proposing a tailored pre-processing or feature engineering method to enhance the model's ability to extract relevant information from the speech signals of the particular target population [3]. This section aims to provide a rational reason for the study, emphasizing the utilization of Deep Neural Networks (DNNs) as a strategic and technologically advanced method to improve Automatic Speech Recognition. The main emphasis is on the potential advantages of Deep Neural Networks (DNNs) in aiding underprivileged communities in the Tamil language. The phases involved in the processing of Speech Dialect are presented in Figure 1.

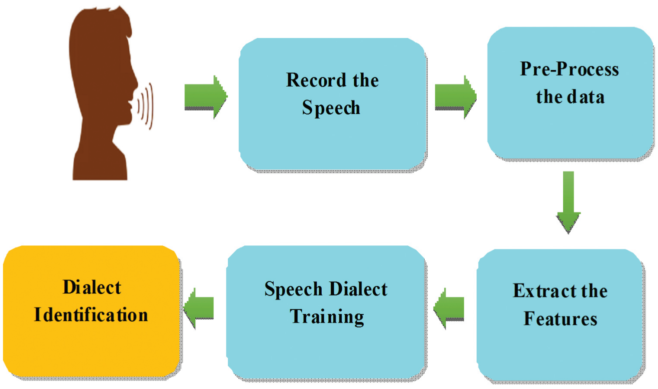


Figure 1: Speech Dialect Processing

2 LITERATURE REVIEW

The authors of [4] explored sub-word dictionary learning and segmentation techniques for Automatic Speech Recognition (ASR) for Tamil and Kannada. The research is extended in [5] by providing robust hearing-impaired speaker recognition from speech, employing deep learning networks in the native language. In the speech recognition model, the authors of [6] implemented a Bidirectional Recurrent Neural Network with a Self-Organizing Map in their research. In contrast, the automation of Tamil speech recognition is highlighted in [7]. In continuation of these references, the design and development of a large vocabulary for a continuous speech recognition system is highlighted in [8]. However, the authors of article [9] contributed their research in finding overlapping speech recognition systems using switching mechanisms between signals, whereas heterogeneous groups of speakers were concentrated in [10].

An interactive system for speech [11] is implemented based on Deep Neural Network (DNN) and i-vector, and performance analysis on syllable-based speech recognition is given in [12]. The comparative study on the models is projected in [13] for the multilingual training process, and [14] focuses on the robust speech recognition based on syllables for which the Grapheme Gaussian model, segmentation [15], and prosodic syllable were focused on by the authors of [16] and monosyllable in [17]. Tamil word recognition using HMM-GMM is given in [18] & [21], along with isolated [19], [22] &[28] and continuous [20] & [24] words with the hybrid model were research performed by various authors. Strategy for spoken word recognition is given in [23] & [25]. Different simulation mechanism for speech recognition is conducted by the researchers, such as Modified Mel Frequency Cepstral Coefficient Algorithm [26], Wavelet Denoising and Hidden Markov Model [27], paralinguistics [29], and Optimal Adaptive Filtering Technique [30] & [31].

An analysis of speech recognition for Tamil and Malay lang using Artificial Neural Network (ANN) [32], voice recognition [33], Automatic Speech Recognition for Tamil and Hindi [34], POS Tagging Using Naïve Bayes Algorithm for Tamil language [35], and part of speech tagging [36] are other mechanisms concentrated by various researchers.

3 PROPOSED METHODOLOGY

3.1 Problem Statement

The project seeks to accomplish multiple objectives focused on accurately identifying and distinguishing regional Tamil dialects by uniquely utilizing Deep Neural Networks (DNNs) with updated features. The main goal is to create a robust DNN acoustic model specifically designed for the complexities of Tamil speech patterns. This research entails employing modified characteristics, such as Mel-Frequency Cepstral Coefficients (MFCC), to amplify the model's responsiveness to the subtleties of other dialects. The study aims to enhance the development of speech-processing technology specifically for Tamil by methodically assessing the effectiveness of the suggested DNN model. The research attempts to discover the most influential parameters that affect the model's efficiency in detecting dialectical changes in the Tamil language through rigorous experimentation. In summary, the study aims to offer valuable insights into the utilization of advanced technologies in the field of dialect identification, explicitly focusing on the intrinsic linguistic diversity seen in Tamil speech.

3.2 Application of DNN and the Implementation of Modified Features

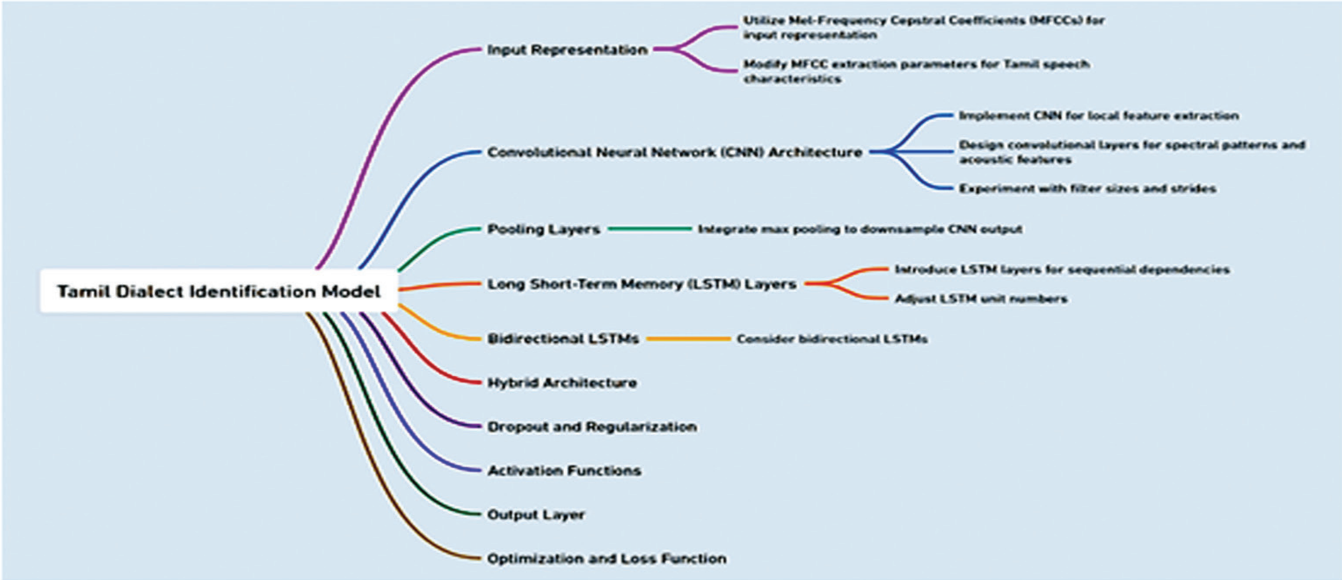


Figure 2: Framework for CNN-LSTM with MFCC in Speech Processing

The research framework proposes the strategic selection of a hybrid CNN-LSTM architecture due to its distinct advantages in effectively addressing the intricate problem of identifying dialects inside the Tamil language. Convolutional Neural Networks (CNNs) excel at catching small-scale spectral patterns, which makes them very suitable for detecting subtle acoustic characteristics seen in speech transmissions. In contrast, Long Short-Term Memory (LSTM) networks excel in capturing sequential dependencies, enabling the model to comprehend the temporal intricacies inherent in spoken language. By integrating these architectures in a hybrid manner, the model acquires expertise in capturing nearby and distant connections. This results in a comprehensive comprehension of the various dialectical differences found in Tamil speech. The complete process of CNN-LSTM with MFCC is projected in Figure 2.

The Mel-Frequency Cepstral Coefficients (MFCCs) are crucial in this hybrid model. The selection of MFCCs as the input feature representation is based on their efficient ability to capture the frequency characteristics of audio signals and is presented in Figure 3. To be precise, extracting MFCC entails decomposing the audio signal into its constituent frequency components, replicating the human auditory system's response to various sound frequencies. The extraction parameters of MFCCs are adapted to better match the distinct tonal qualities and pronunciation nuances present in Tamil speech. This sophisticated portrayal acts as a polished input for the hybrid CNN-LSTM model, augmenting its ability to discern dialectical changes.

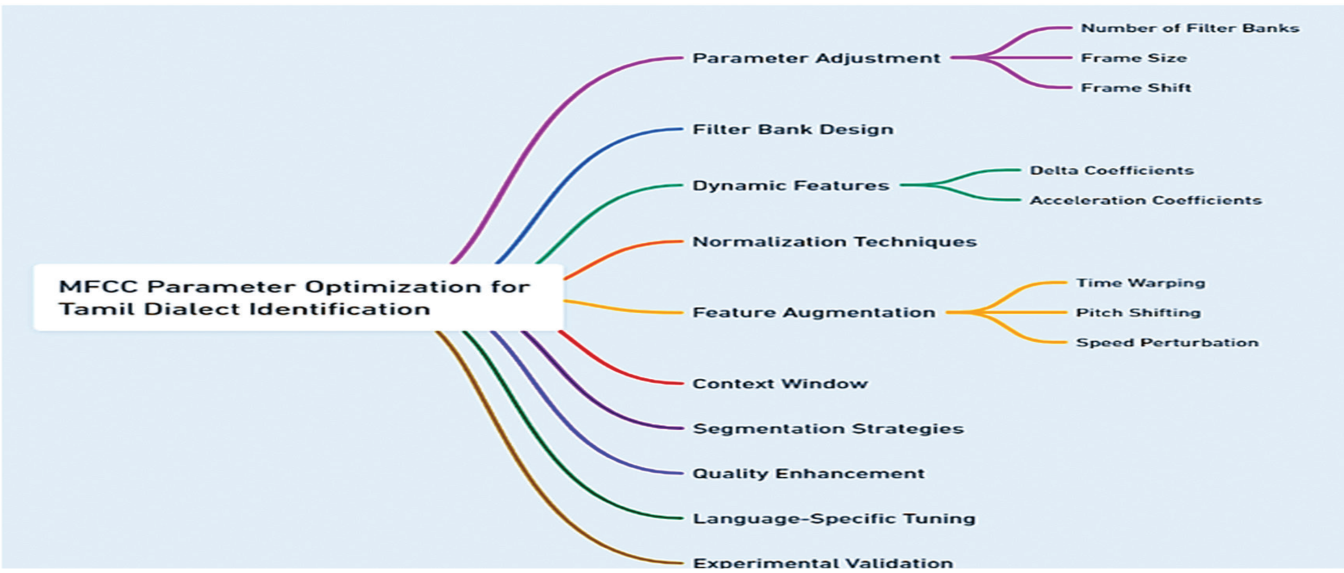


Figure 3: MFCC Parameter Optimization

The hybrid CNN-LSTM architecture is selected to exploit the synergistic advantages of CNNs and LSTMs, resulting in a model that can effectively capture local and sequential data essential for dialect recognition in Tamil. By including MFCCs, this strategy enhances the model's input representation to match Tamil speech's unique features better. As a result, the model's performance is optimized to detect and distinguish regional dialects within the language reliably.

4 SPEECH IDENTIFICATION MODEL DESIGN

The Speech Identification Design model utilizes a hybrid architecture consisting of Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs), especially a CNN-LSTM model, to perform the task of voice identification. The motivation behind this approach is to efficiently capture the spatial and temporal correlations found in the intricate patterns of voice signals. The CNN module demonstrates exceptional proficiency in extracting hierarchical spatial features from the Mel-Frequency Cepstral Coefficients (MFCC) representations of speech. This capability allows the model to identify and distinguish significant auditory patterns accurately. The LSTM component simultaneously deals with the temporal dynamics present in speech, enabling the comprehension of the sequential characteristics of phonetic and language aspects. The MFCC features are customized to suit the specific attributes of the Tamil language, guaranteeing that the model is proficient at detecting the subtleties and deviations in spoken languages. These alterations may involve tailored preprocessing procedures or precise adjustments to the feature representation, improving the model's ability to identify Tamil speech accurately. The combination of CNN-LSTM architecture and modified MFCC features demonstrates a holistic approach to speech processing, highlighting the collaboration between sophisticated neural network structures and domain-specific feature engineering to enhance performance in the field of Tamil speech identification.

Algorithm Name: Tamil Speech Identification

1. **Method: Data Collection and Preprocessing:**
    a. Gather a varied dataset of Tamil voice samples.
    b. Preprocess the raw audio data
        i. extracting the Mel-Frequency Cepstral Coefficients (MFCCs).
        ii. Implement any required modifications.
2. **Methodology: Feature Engineering and Modification**
    a. Apply alterations to the MFCC features:
        i. Preprocessing processes tailored to specific requirements.
        ii. Modifications to the portrayal of features.
        iii. Integrate linguistic and cultural knowledge.
3. **Designing the Procedure Model Architecture:**
    a. Specify the architecture of the CNN-LSTM model:
        i. Please provide the exact number and arrangement of layers.
        ii. Activation functions are mathematical functions that introduce non-linearity into neural networks. They are applied to the output of each neuron in a network to determine whether it should be activated based on a certain threshold. Activation functions help neural networks learn complex patterns.
        iii. Combine spatial characteristics derived by Convolutional Neural Networks (CNNs) with temporal dependencies using Long Short-Term Memory (LSTM) models.
4. **The Train Model procedure involves**
    a. dividing the dataset into separate training and validation sets.
    b. Conduct training for the CNN-LSTM model:
        i. Optimize parameters using the backpropagation algorithm.
        ii. Employ gradient descent to minimize the loss function.
5. **The Validate And Tune Hyperparameters procedure:**
    a Verify the accuracy of the trained model using a distinct dataset.
    b. Adjust hyperparameters if needed:
        i. Enhance efficiency and mitigate overfitting.
6. **Performance of the Procedure Evaluation:**
    a. Assess the model's performance using metrics:
        i. Accuracy, precision, recall, F1-score, and dialect-specific metrics.
        ii. Examine confusion matrices.
7. **Method of Interpreting and Analyzing:**
    a. Perform interpretability analysis:

i. Utilize visualizations or conduct feature significance analysis.

ii. Comprehend the specific traits that the model prioritizes for dialect recognition.

8. The User Feedback And Iterative Improvement procedure involves

a. collecting feedback from users, mainly from specific target demographics.

b. Utilize feedback to enhance the model progressively:

i. Resolve identified issues or address raised concerns.

9. The Deployment and Integration Procedure involves:

a. the implementation of the trained model into a practical application or system.

b. Incorporate the model into the intended setting.

10. Procedure Continuous Monitoring and Maintenance:

a. Establish mechanisms for ongoing monitoring of model performance.

b. Continuously refresh the model by incorporating new data and insights.

11. Terminate the algorithm

Mathematical Modeling of the Proposed Methodology

1. MFCC Extraction

a. Pre-emphasis:

$$x'(t) = x(t) - \alpha \cdot x(t-1) \tag{1}$$

b. Frame Blocking:

$$X_n(t) = \sum_{n=1}^{N} w(t) \cdot \frac{d}{dt} x'(t - n \cdot \text{frame shift}) \tag{2}$$

c. Fourier Transform:

$$X_n(f) = \int_{-\infty}^{\infty} x_n(t) \cdot e^{-j2\pi f t} dt \tag{3}$$

d. Mel Filterbank:

$$S_m - \sum_{k-1}^{K} H_m(k) \cdot |X_n(k)|^2 \tag{4}$$

e. Logarithm:

$$\text{MFCC}_m = \log\left(\int_{-\infty}^{\infty} S_m(t) dt\right) \tag{5}$$

2. CNN for Feature Extraction

$$F_{i,j} = \sigma\left(\sum_{m-1}^{M} \sum_{n-1}^{N} \int_{-\infty}^{\infty} x_{i+m,j-n}(t) \cdot W_{m,ir}(t) dt + b\right) \tag{6}$$

3. Temporal Modeling with LSTM

$$f_t = \int_{-\infty}^{\infty} \sigma\big(W_{xf}(t) \cdot x_t(t) + W_{hf}(t) \cdot h_{t-1}(t) + b_f\big) dt \tag{7}$$

$$i_t = \int_{-\infty}^{\infty} \sigma\big(W_{xi}(t) \cdot x_t(t) + W_{h_i}(t) \cdot h_{t-1}(t) + b_i\big) dt \tag{8}$$

$$\bar{c}_t = \int_{-\infty}^{\infty} \tanh\left(W_{xc}(t) \cdot x_t(t) + W_{hc}(t) \cdot h_{t-1}(t) + b_c\right) dt \tag{9}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \vec{c}_t \tag{10}$$

$$o_t = \int_{-\infty}^{\infty} \sigma\left(W_{xo}(t) \cdot x_t(t) + W_{ho}(t) \cdot h_{t-1}(t) + b_o\right) dt \tag{11}$$

$$h_{t_1} = o_t \cdot \int_{-\infty}^{\infty} \tanh\left(c_i(t)\right) dt \tag{12}$$

4. Model Output

$$y - \text{Softmax}\left(\int_{-\infty}^{\infty} W_y(t) \cdot h_t(t) dt + b_y\right) \tag{13}$$

5. Loss Function

$$L = -\sum_i \int_{-\infty}^{\infty} y_i(t) \cdot \frac{d}{dt} \log\left(\hat{y}_i(t)\right) dt \tag{14}$$

6. Optimization

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \int_{-\infty}^{\infty} \nabla_0 L dt \tag{15}$$

7. Feature Enhancement

$$_i(t) = \int_{-\infty}^{\infty} f_i(x(t)) \cdot \frac{d}{dt} x(t) dt \tag{16}$$

8. Temporal Integration in CNN

$$F_{i,j}(t) = \sigma\left(\sum_{m+1}^{M} \sum_{\pi^2+1}^{N} \int_{-\infty}^{\infty} x_{i-m,j-m}(t') \cdot W_{m,i}(t-t') dt' + b\right) \tag{17}$$

9. Dynamic LSTM

$$f_t = \int_{-\infty}^{\infty} \sigma\big(W_{ef}(t-t') \cdot x_t(t') + W_{hf}(t-t') \cdot h_{t-1}(t') + b_f\big) dt' \tag{18}$$

$$i_t - \int_{-\infty}^{\infty} \sigma\big(W_{\pi i}(t-t') \cdot x_t(t') + W_{h_i}(t-t') \cdot h_{t-1}(t') + b_i\big) dt' \tag{19}$$

$$\varepsilon_t - \int_{x_0}^{\infty} \tanh\left(W_{xc}(t-t') \cdot x + (t') + W_{hc}(t-t') \cdot h_t h_{i1}(t') + b_c\right) dt' \tag{20}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t \tag{21}$$

$$o_t = \int_{-\infty}^{\infty} \sigma\big(W_{xo}(t-t') \cdot x_t(t') + W_{h_0}(t+t') \cdot h_{t-1}(t') + b_0\big) dt' \tag{22}$$

$$h_t = o_t \cdot \int_{-\infty}^{\infty} \tanh\big(c_t(t')\big) dt' \tag{23}$$

10. Temporal Integration in Model Output

$$y(t) = \text{Softmax}\left(\int_{-\infty}^{\infty} W_y(t-t') \cdot h_t(t')dt' + b_y\right) \quad (24)$$

11. Continuous Loss Function

$$L(t) = -\sum_i \int_{-\infty}^{\infty} y_i(t') \cdot \frac{d}{dt}\log\left(\hat{y}_i(t')\right)dt' \quad (25)$$

12. Continuous Optimization

$$\theta_{\text{new}}(t) = \theta_{\text{old}} - \eta \cdot \int_{-\infty}^{\infty} \nabla_\theta L(t')dt' \quad (26)$$

5. EXPERIMENTAL SETUP FOR TAMIL DIALECT RECOGNITION

The Tamil Speech Dialect dataset is taken from the site [37]. The dataset comprises accurately transcribed audio recordings of Tamil phrases, which were recorded by volunteers and are of excellent quality. The dataset consists of wave files and a TSV file named line_index.tsv. The line_index.tsv file contains a FileID that has been anonymized, as well as audio transcription within the file. The dataset is categorized into male and female recordings by making 1956 recordings for the prior and 2335 recordings for the latter, respectively. Total recordings count to 4291 with 1.5GB of data. In this research, a random sampling of 250 recordings in each category to 500 is considered for analysis. Among these 500 recordings, 400 are considered for training and the remaining 100 for testing purposes, making a ratio of 80:20. The experiment for Tamil Dialect Recognition is implemented in Python using Google CoLab by loading the data in the Google Drive.

6. RESULTS AND DISCUSSION

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 32)	832
dropout_1 (Dropout)	(None, 32)	0
dense_2 (Dense)	(None, 16)	528
dropout_2 (Dropout)	(None, 16)	0
dense_3 (Dense)	(None, 4)	68

Total params: 1,428
Trainable params: 1,428
Non-trainable params: 0

Figure 4: Number of Parameters used for dialect recognition using CNN-LSTM-MFCC

The above Figure 4 represents the number of parameters used in the recognition of dialect with the implementation of the CNN-LSTM-MFCC model is projected.

```
Epoch 1/50
271/271 [==============================] - 0s 500us/step - loss: 1.1195 - acc: 0.4649 - val_loss: 0.7927 - val_acc: 0.6912
Epoch 2/50
271/271 [==============================] - 0s 102us/step - loss: 0.6926 - acc: 0.7011 - val_loss: 0.6229 - val_acc: 0.7353
Epoch 3/50
271/271 [==============================] - 0s 129us/step - loss: 0.6038 - acc: 0.7417 - val_loss: 0.6628 - val_acc: 0.7206
Epoch 4/50
271/271 [==============================] - 0s 116us/step - loss: 0.5601 - acc: 0.7749 - val_loss: 0.6381 - val_acc: 0.7206
Epoch 5/50
271/271 [==============================] - 0s 138us/step - loss: 0.5535 - acc: 0.7712 - val_loss: 0.6085 - val_acc: 0.7206
Epoch 6/50
271/271 [==============================] - 0s 104us/step - loss: 0.4949 - acc: 0.8118 - val_loss: 0.6371 - val_acc: 0.7500
Epoch 7/50
271/271 [==============================] - 0s 121us/step - loss: 0.5626 - acc: 0.7860 - val_loss: 0.6192 - val_acc: 0.7500
Epoch 8/50
271/271 [==============================] - 0s 115us/step - loss: 0.4464 - acc: 0.8081 - val_loss: 0.6022 - val_acc: 0.7206
Epoch 9/50
271/271 [==============================] - 0s 116us/step - loss: 0.4946 - acc: 0.8155 - val_loss: 0.6074 - val_acc: 0.7647
Epoch 10/50
271/271 [==============================] - 0s 139us/step - loss: 0.4280 - acc: 0.8044 - val_loss: 0.6904 - val_acc: 0.7206
Epoch 11/50
271/271 [==============================] - 0s 101us/step - loss: 0.4828 - acc: 0.8266 - val_loss: 0.7194 - val_acc: 0.7059
Epoch 12/50
271/271 [==============================] - 0s 99us/step - loss: 0.4583 - acc: 0.7934 - val_loss: 0.5883 - val_acc: 0.7500
Epoch 13/50
271/271 [==============================] - ETA: 0s - loss: 0.6956 - acc: 0.600 - 0s 114us/step - loss: 0.4146 - acc: 0.8192 - v
al_loss: 0.6043 - val_acc: 0.7500
Epoch 14/50
271/271 [==============================] - 0s 131us/step - loss: 0.6409 - acc: 0.7638 - val_loss: 0.7513 - val_acc: 0.5735
Epoch 15/50
271/271 [==============================] - 0s 133us/step - loss: 0.4661 - acc: 0.8155 - val_loss: 0.6691 - val_acc: 0.7500
```

Figure 5: Accuracy Calculation using CNN-LSTM-MFCC

The Figure 5 represents the output of the implementation by highlighting the accuracy parameter for the proposed model.

Table 1: Performance Analysis on Accuracy

S.No	Dialects	No. Of Training Samples	No. of Test Samples Correctly Recognized by Models			Accuracy		
			HMM	GMM	CNN-LSTM-MFCC	HMM	GMM	CNN-LSTM-MFCC
1	Northern Tamil	133	115	120	130	86	90	97
2	Southern Tamil	142	127	118	138	89	83	97
3	Eastern Tamil	125	103	112	115	82	89	92
	TOTAL	400	345	350	383	86	87	95

7. CONCLUSION AND FUTURE ENHANCEMENTS

This study represents a significant advance in the field of Tamil speech recognition. By harnessing the power of DNN and incorporating carefully designed features, the study aims to improve the accuracy and efficiency of speech recognition adapted to the nuances of the Tamil language. Integrating neural networks, especially hybrid architectures such as CNN-LSTM, emphasizes the commitment to capture both the spatial and temporal complexities inherent in speech signals. The feature modification, possibly including custom Mel-Frequency Cepstral Coefficients (MFCC) modifications, speaks of the commitment to addressing the unique characteristics of Tamil dialects. This work aligns with broader trends in speech recognition research, where advanced neural network architectures and nuanced feature design play a key role in achieving state-of-the-art results. The future work might include more comprehensive research on speech recognition and cover various approaches, including robust recognition of hearing-impaired speakers, sub-word learning, and analysis of Tamil language speech recognition systems using multiple techniques and models will be focused.

REFERENCES

1. Varsha Balaji, Archana Jp, and Bharathi B. 2023. Automatic Speech Recognition vulnerable old-aged and transgender people in Tamil. LTEDI.
2. Sri Harish, P. Vijayalakshmi, and T. Nagarajan. 2011. Significance of segmentation in phoneme-based Tamil speech recognition system. International Conference on Electronic Computer Technology DOI:https://doi.org/10.1109/icectech.2011.5941739
3. Kavitha Vijayakumar, Saranyaa Gunalan, and Ranjith Rajeshwaran. 2021. Development of Minimal Pair Test in Tamil (MPT-T). Journal of Clinical and Diagnostic Research. DOI: https://doi.org/10.7860/jcdr/2021/46807.15357
4. Madhavaraj A, Bharathi Pilar, Ramakrishnan A G, and A. G. Ramakrishnan. 2022. Subword Dictionary Learning and Segmentation Techniques for Automatic Speech Recognition in Tamil and Kannada. DOI:https://doi.org/10.48550/arxiv.2207.13331
5. KiranBala Benny and Viswanathan Balasubramanian. 2023. Robust Hearing-Impaired Speaker Recognition from Speech using Deep Learning Networks in Native Language. The international Arab journal of information technology. DOI:https://doi.org/10.34028/iajit/20/1/11
6. Priyan Malarvizhi Kumar. 2022. Retraction Note: An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self-Organizing Map. Neural computing & applications. DOI:https://doi.org/10.1007/s00521-022-08144-x
7. S. Lokesh, Priyan Malarvizhi Kumar, M. Ramya Devi, P. Parthasarathy, and Chandra Babu Gokulnath. 2019. An Automatic Tamil Speech Recognition system by using Bidirectional Recurrent Neural Network with Self-Organizing Map. Neural Computing and Applications. DOI:https://doi.org/10.1007/s00521-018-3466-5
8. Madhavaraj, and A. G. Ramakrishnan. 2017. Design and development of a large vocabulary, continuous speech recognition system for Tamil. IEEE India Conference. DOI:https://doi.org/10.1109/indicon.2017.8488025
9. Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, K. Kinoshita, Naoyuki Kamo, and Takafumi Moriya. 2022. Learning to Enhance or Not: Neural Network-Based Switching of Enhanced and Observed Signals for Overlapping Speech Recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing. DOI:https://doi.org/10.1109/icassp43922.2022.9746347
10. Romain Serizel, and Diego Giuliani. 2017. Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children. Natural Language Engineering. DOI:https://doi.org/10.1017/s135132491600005x
11. P. Shanmugapriya, V. Mohan, S. Yogapriya, and Y. Venkataramani. 2018. Speech Based Interaction System Using DNN and i-vector. International Conference on Recent Trends in Image Processing and Pattern Recognition. DOI:https://doi.org/10.1007/978-981-13-9181-1_41
12. S. Sundarapandiyan, Shanthi, and M. Yoonus. 2016. Performance Analysis of Syllable Based Tamil Language Robust Speech Recognition System Using Modified Group Delay Function, Gammatone Cepstral Coefficients, Hidden Markov Model and Deep Neural Network.
13. Haihua Xu, Van Hai Do, Xiong Xiao, and Eng Siong Chng. 2015. A comparative study of BNF and DNN multilingual training on cross-lingual low-resource speech recognition. Interspeech. DOI:https://doi.org/10.21437/interspeech.2015-481
14. M. Yoonus. 2016. Syllable Based Tamil Language Continuous Robust Speech Recognition Using MGDF.
15. Akila, and Chandra E. 2015. Word Based Tamil Speech Recognition Using Temporal Feature Based Segmentation. ICTACT Journal on Image and Video Processing. DOI:https://doi.org/10.21917/ijivp.2015.0152
16. Akila A. Ganesh, and Chandra Ravichandran. 2013. Grapheme Gaussian model and prosodic syllable-based Tamil speech recognition system. International Computer Science Conference. DOI:https://doi.org/10.1109/icspcom.2013.6719821
17. Geetha K, and Chandra E. 2015. Monosyllable Isolated Word Recognition for Tamil language using Continuous Density Hidden Markov Model. IEEE International Conference on Electrical, Computer and Communication Technologies. DOI:https://doi.org/10.1109/icecct.2015.7226056
18. Geetha K, and Vadivel R. 2017. Estimation of HMM-GMM Parameter for Tamil Words. International journal of engineering and technology. DOI:https://doi.org/10.21817/ijet/2017/v9i1/170901412
19. Janani J. and Mohanapriya N. 2013. Isolated Tamil Speech Recognition System Based On Cmu Sphinx. International Journal of Information Technology & Computer Sciences Perspectives.
20. M. Kalamani, S. Valarmathy, and M. Krishnamoorthi. 2015. Hybrid Modeling Algorithm for Continuous Tamil Speech Recognition.
21. S. Karpagavalli, and E. Chandra. 2015. Phoneme and word based model for Tamil speech recognition using GMM-HMM. 2015 International Conference on Advanced Computing and Communication Systems. DOI:https://doi.org/10.1109/icaccs.2015.7324119
22. J. Murali Krishna, and M. Vanitha Lakshmi. 2014. Speaker Independent Isolated Tamil Words for Speech Recognition using MFCC, IPS and HMM.
23. S. Palanivel. 2012. Spoken Word Recognition Strategy for Tamil Language.
24. V. Radha, C. Vimala, and M. Krishnaveni. 2012. Continuous Speech Recognition system for Tamil language using monophone-based Hidden Markov Model. International Conference on Computational Science, Engineering and Information

Technology. DOI:https://doi.org/10.1145/2393216.2393255

25. S. Rojathai. 2018. Spoken Tamil word Recognition System.

26. Swagata Sarkar, Sanjana R, Rajalakshmi S, and Harini T J. 2018. Simulation and detection Of Tamil Speech Accent Using Modified Mel Frequency Cepstral Coefficient Algorithm. International journal of engineering and technology. DOI:https://doi.org/10.14419/ijet.v7i2.33.14202

27. C. Vimala, and V. Radha. 2013. Efficient Speaker Independent Isolated Speech Recognition for Tamil Language Using Wavelet Denoising and Hidden Markov Model. DOI:https://doi.org/10.1007/978-81-322-1000-9_52

28. K. Yogendirakumar, A. Weerasinghe, and W. G. D. M. Wathugala. 2004. Isolated-Word Speech Recognition for Tamil Language using Hidden Markov Models.

29. Anosha Ignatius and Uthayasanker Thayasivam. 2021. A Survey on Paralinguistics in Tamil Speech Processing. Dravidianlangtech.

30. V. Jagannaveen, T. Prabakar, J. V. Suman, P. Devi, Pradeep, QuotNoise, J. Vanus, P. Loizou, K. Borisagar, S. Prabha, and S. Nandyala. 2016. Optimal Adaptive Filtering Technique for Tamil Speech Enhancement.

31. Kingston Pal Thamburaj, Ponniah K, Kartheges Ponniah, Ilangkumaran Sivanathan, Manoj Kumar, and Muniisvaran Kumar. 2021. An Critical Analysis of Speech Recognition of Tamil and Malay Language Through Artificial Neural Network. DOI:https://doi.org/10.20944/preprints202102.0156.v1

32. R Kiran, K Nivedha, S Pavithra Devi, and T Subha. 2017. Voice and speech recognition in Tamil language. International Conference on Computing and Convergence Technology. DOI:https://doi.org/10.1109/iccct2.2017.7972293

33. Rahul Mishra, Senthil Raja Gunaseela Boopathy, Manikandan Ravikiran, Shreyas Kulkarni, Mayurakshi Mukherjee, Ananth Ganesh, and Kingshuk Banerjee. 2023. Revisiting Automatic Speech Recognition for Tamil and Hindi Connected Number Recognition. DRAVIDIANLANGTECH.

34. Radha. 2012. Optimal Adaptive Filtering Technique for Tamil Speech Enhancement. International Journal of Computer Applications. DOI:https://doi.org/10.5120/5633-7996

35. Rajasekar and A. Udhayakumar. 2020. POS Tagging Using Naïve Bayes Algorithm For Tamil. International Journal of Scientific & Technology Research.

36. Hemakasiny Visuwalingam, Ratnasingam Sakuntharaj, and Roshan G. Ragel. 2021. Part of Speech Tagging for Tamil Language Using Deep Learning. International Conference on Industrial and Information Systems. DOI:https://doi.org/10.1109/iciis53135.2021.9660738

37. Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association (ELRA), Marseille, France, 6494–6503. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.800

SENTIMENT ANALYSIS AND EMOTION RECOGNITION IN TAMIL TEXT & SPEECH

Sentiment analysis on Electoral data using Adapter fusion, a multi-task learning approaches

Vijay Sundar Ram R, Pattabhi RK Rao, Sobha Lalitha Devi

ABSTRACT

Social media platforms such as Facebook, Twitter etc have turned out to be the channels for people of all levels to register their opinions, feelings, feedback on various events happening around them. This information is very essential for understanding the sentiments of wide-range of people on various events. It is useful in various aspects such as product marketing, behaviour analysis, strategic planning and pandemic management etc. Political social media data provides a strong people's sentiment on different parties, leaders, schemes which has become a vital input in pre-poll analysis. In this paper, we present our work on sentiment analysis on Tamil electoral data, collected from twitter. We use a multi-task learning approach, Adapter Fusion. In our work, first task is to identify the electoral entities. We consider the named entities such as party names, political leaders, place names, schemes, election manifesto etc as electoral entities. Second task is to determine the sentiments related to these entities. We learn task specific parameters called adapters that encapsulate the task-specific information and then combine the adapters in a separate knowledge composition step to identify the electoral entities and sentiments related to those entities. The results are encouraging.

Vijay Sundar Ram R, Pattabhi RK Rao, Sobha Lalitha Devi
AU-KBC Research Centre, MIT Campus of Anna University, Chennai.
{sobha, sundar, pattabhi}@au-kbc.org

1 INTRODUCTION

Social media has become the promising source of communication across all levels, as millions of users share their opinions about the products, celebrities, movies, and politics in the social media sites such as twitter, face-book and other discussion forums. With availability of various native language inputting tools, people prefer to communicate in their mother tongue and share their opinions. This has led to a rise of tweets in Indian languages. There is a great demand from business and commercial perspective, to extract potential information from these unstructured data. This information is very essential for understanding the sentiments of wide-range of people on various events. It is useful in various aspects such as product marketing, behaviour analysis, strategic planning and pandemic management etc.

Political social media data provides a strong people's sentiment on different parties, leaders, schemes which has become a vital input in pre-poll analysis. In this paper, we present our work on sentiment analysis on Tamil electoral data, collected from twitter. We use a multi-task learning approach, Adapter Fusion.

For analysing the sentiments, earlier the systems were developed using different methodologies, namely, Lexicon based approach, Machine Learning based approach, which includes classifiers such as Decision tree, k-nearest neighbour, probabilistic classifiers, and linear classifiers and Hybrid approach.

With successful machine translation systems using RNN, Deep learning approaches was used to build Sentiment analysis systems. Wang et. Al. (2016.a) presented a Sentiment analysis system for short text using the combination of Conventional Neural Networks (CNN) and Recurrent Neural Networks (RNN), where the author has claimed that course-grained local features were generated by CNN and long-distance dependencies are learned using RNN. But learning long-distance dependencies was a challenge. Wang et. al. (2016.b) used CNN-LSTM (Long Short Term Memory) where LSTM helped in learning long-distance dependencies. Zhang et. al. (2018) came up with a details survey on sentiment analysis work done using various Deep learning techniques such as CNN, RNN, LSTM, Attention mechanism with RNN.

After the development of BERT (Bidirectional Encoder Representation from Transformers) by Devlin et. Al. (2018), researchers started to use it in developing and fine tuning the sentiment analysis systems. A version of BERT, a un-normalised multilingual model, contains 104 languages. Karini and Sharabadi (2019) published a work on sentiment analysis in Persian text using multilingual BERT with attention model. Visser and Dunaiski (2022) used various BERT models for classifying the intent and sentiment classification of in-text citations of articles in ACM library. They found BERT-cased and Sci-BERT-cased model to perform best.

With the availability of Large Language Models (LLMs) recent publications are on sentiment analysis using LLM. Kheiri and Karimi (2023) have published SentimentGPT, where they have presented an experiment using different GPT models evaluated on SemEval 2017 dataset. Sun et. al. (2023) has published a sentiment analysis work using multiple LLMs, where the complementary abilities of different models helped in improving the performance.

Though an exhaustive work on Sentiment Analysis has been published in English, Chinese and few European languages, there are very less works published on Sentiment Analysis in Tamil. Thavarseen and Mahesun (2019) presented a sentiment analysis in Tamil using K-means and K-nodes with Bag of Words (BoG) techniques. Sharmistra (2020) has done her doctoral dissertations in sentiment analysis for products in Tamil reviews available on social media using different classifier based machine learning algorithms. The shared task, Sentiment Analysis in Tamil and Tulu- DravidianLangTech@RANLP 2023 conducted by Hedge et al (2023) has given a boost to sentiment analysis research in Tamil. Shanmugavadivel et al (2022) has presented sentiment analysis in Tamil code-mixed data using CNN+BiLSTM.

We present our work on sentiment analysis in electoral data collected from Tamil twitter data. Here the sentiments are related to different entities such as person, parties, schemes etc. Thus we need to identify entity based sentiment. To achieve this we need to find the entities followed by identifying the sentiments related to the entities. We propose to build an end-to-end system for sentiment analysis using multi task transfer learning algorithm with Electoral Entity Recognition and Sentimental analysis as two sub tasks. We choose to use Adapter Fusion-based multi task learning methodology proposed by (Pfeiffer et al, 2021), which is a non-destructive multi-task learning algorithm.

Further the paper is organised as follows: In the following section we have presented a short notes on different multi-task learning approaches, Adapter Fusion techniques. Section 3 has the description of our experiments. Data set is explained in section 4. Section 5 has the results and discussion. The paper ends with the conclusion and Limitation sections.

2 MULTI-TASK LEARNING

For developing an end-to-end Sentiment analysis system, we used AdapterFusion-based multi-task learning methodology (Pfeiffer et al, 2021). This transfer learning method has two stage learning, knowledge extraction stage and knowledge composition step. A brief description on transfer learning, Adapter and Adapterfusion are given below.

The two predominant approaches in transfer learning for sharing knowledge across multi-task are

- **Sequential Fine-tuning:**

This involves sequentially updating all the weights of the model on each task. This approach works well for two sequential tasks and beyond that leads to catastrophic forgetting.

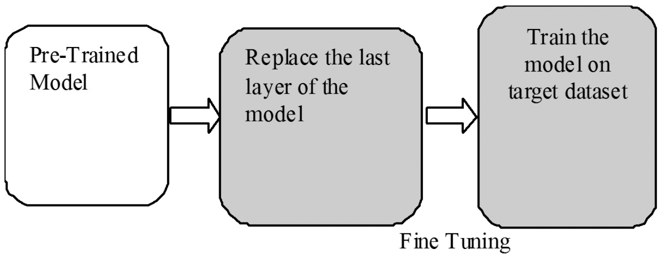


Figure 1: Sequential Fine-tuning

- Multi-Task Learning (MTL):

All tasks are trained simultaneously with the aim of learning a shared representation that will enable the model to generalize better on each task. More tasks cannot be added as MTL requires to simultaneously access all the tasks.

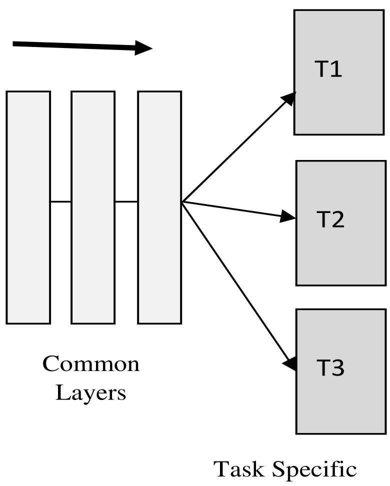


Figure 2: Multi-Task Learning

2.1 Adapters

To overcome the limitation in Sequential Fine-tuning and MTL, Houlsby et al (2019) introduced adapters, which do not require fine-tuning of all the parameters of the pre-trained model and instead introduce a small number of task specific parameters while keeping the underlying pre-trained language model fixed. Adapters share a large set of parameters Θ across all tasks and introduce a small number of task-specific parameters Φn. While Θ represents the weights of a pre-trained model (e.g., a transformer), the parameters Φn, where n □ {1, . . . , N}, are used to encode task-specific representations in intermediate layers of the shared model. There are two variants of adapters, namely Single task Adapter, where different Adapters are trained for each of the N task and Multiple Task Adapter, where N task is trained in parallel (Stickland and Murray, 2019).

2.2 Adapter Fusion

To maximize the transfer of knowledge across tasks, without suffering from catastrophic forgetting and difficulties in dataset balancing, AdapterFusion was introduced by Pfeiffer et al (2021). After the training of the task specific Adapters, these Adapter are combined using AdapterFusion. Once training for the adaptersΦm and again for training Fusion parameters Ψm, which learn to compose the information stored in the N task adapters. It learns to compose the N task adapters Φn and the shared pre-trained model Θ, by introducing a new set of weights Ψ. These parameters learn to combine the adapters as a dynamic function of the target task data.

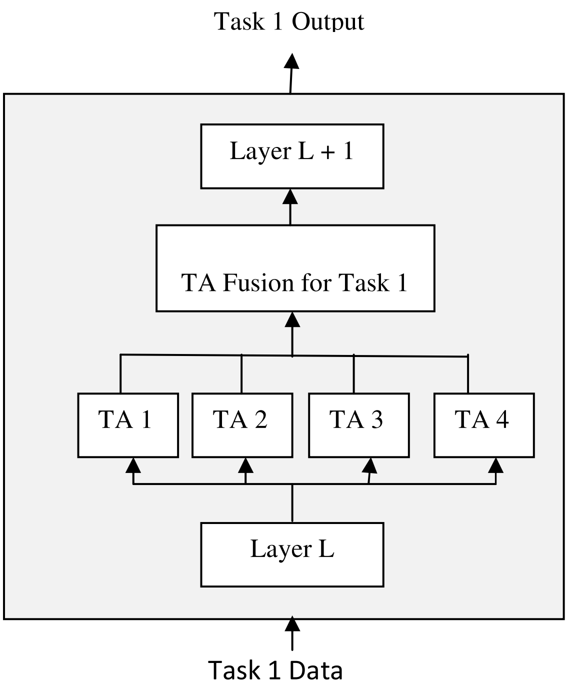


Figure 1: AdapterFusion

3. OUR APPROACH

End-to-end sentiment analysis requires the following task, Electoral entity recognizer and sentiment analysis system. We train each task as a single Task Adapter. The adapters are combined in the AdapterFusion task and Fusion parameters Ψ is learned.

3.1. Experimental Setup

In training all the four adapters, we use XLM-R as the pre-trained language model. We train them with reduction factors {2, 16, 64} and learning rate 0.0001 with AdamW and a linear learning rate decay. We train for a maximum of 30 epochs with early stopping as used by (Pfeiffer et al, 2021). For AdapterFusion, We used a learning rate of 5e−5. We trained for a maximum of 10 epochs with early stopping. While we initialize Q and K randomly, we initialize V with a diagonal of ones and the rest of the matrix with random weights having a small norm (1e − 6) as mentioned by Pfeiffer et al, (2021).

4. DATASET

We collected electoral data from twitter. This was periodically executed using the publically available twitter APIs. We collected Tamil tweets using latitude-longitude information and using specific keywords such as person, organisation names, events, region specific etc. The twitter data has intrinsic challenges when used in Natural Language Processing applications. The challenges are listed below.

- Less contextual information as the tweets size has restrictions. This leads to more ambiguous tweets.
- As other social media text, it also free flow writing style where there are many repetition of words such as ‘soo so good’.
- They do not have proper punctuation markers.
- Tweets have partial entities and have many spelling variations and mistakes.
- Code-mixing, spoken form and dialectal variations are common in tweets.

These characteristics of tweets make the classification of entities and extraction of information a challenging task.

We periodically collected tweets for three months prior to the elections. The statistics of the corpus is given in table 1 below.

Description	Tamil
No. of tweets	135,000

Table1: Corpus Statistics

For the present task, we have annotated electoral entities, and sentiment in the tweets. We used a GUI enabled tool, PALinkA (Orasen, 2003), an open source tool from University of Wolverhampton. The statistics of named entities and the sentiment tagged data in given in the table 2.

5. RESULTS AND DISCUSSION

We divided the data randomly into 80-20. 80% of the documents were used for training and the 20% is used for testing. Tamil is a morphologically rich and highly agglutinative. We pre-processed the data by tokenizing the agglutinated words into separate words and morphologically segmenting the inflected words.

S.No	Type	Number of Occurrence Tamil
1	Named Entities	125457
2	Positive Tweets	62758
3	Negative Tweets	42745

Table 2: Distribution of Named entities and the sentiment tagged tweets

So we planned to present the data in two different tokenizations. Thus we have two types of experiments using the two different data sets. The different tokenisation's are as follows:

- data set only with words
- Morph-level where word is morphologically analysed and separated as root word and suffixes.

The performance measures for these two experiments for Electoral Entity recognizer (EER), is presented in table 3. Sentiment Analysis performance scores for the tweets are presented in table 4.

S No	Exp	EER		
		Precision%	Recall%	F1%
1	Adapter Fusion word-level	73.34	61.43	66.86
2	Adapter Fusion morph-level	80.55	68.24	73.88

Table 3: Evaluation of EER

Table 3 and Table 4 show the performance scores for Electoral Entity Recognizer and sentiment analysis on electoral data.

As mentioned in section 4, the less contextual information, code-mixing in tweets affect the identification in both named entity reorganization and sentiment analysis. Other characters such as use of pronouns, implicitly writing the information and use of sarcastic writing style pose challenge in sentiment analysis. Filtering of junk characters/non standard characters, short urls and emoticon is a challenging task due to non-standard usages. This depreciates the performance scores. Many long ungrammatical sentences are used in tweets. These make the task tougher.

S. No	Type	Performance Measure	Adapter Fusion word-level	Adapter Fusion morph-level
1	Positive	Pr	65.34	70.34
		Re	78.45	80.23
		F1	71.29	74.96
2	Negative	Pr	67.36	72.46
		Re	75.44	79.56
		F1	71.17	75.84
3	Recall	Pr	58.23	62.56
		Re	74.34	76.45
		F1	65.30	68.81

Table 4: Evaluation of Sentiment Analysis Engine

Thus we can improve sentiment analysis by handling emoticons and sarcasm, including pronominal resolution, and effective filtering the non-standard/ junk characters and short urls. Though the precision is comparatively low, this is very helpful in predicting the trends.

6. CONCLUSION

In this paper, we present an end to end architecture for sentiment analysis of electoral data using AdapterFusion, a new two stage learning algorithm that leverages knowledge from multiple tasks. First task is in identifying the named entities in the text and the second to analysis the sentient in the tweets. We evaluated using periodically collected data from twitter using publically available twitter crawling APIs. The results obtained were very helpful in predicting the trend.

Limitation

Pronominal usage and sarcasm are not handled. Implicit information and exospheric reference are not handled. Non standard acronyms and spelling errors are not handled.

REFERENCE

- Hegde Asha and G Kavya and Coelho Sharal and Lamani Pooja and Shashirekha Hosahalli Lakshmaiah and Chakravarthi, Bharathi R. and Priyadharshini, Ruba and M, Anand Kumar and Thavareesan, Sajeetha and Sherly, Elizabeth, 2023,MUNLP@ DravidianLangTech2023: Learning Approaches for Sentiment Analysis in Code-mixed Tamil and Tulu Text, In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, RANLP 2023, pp 275--281
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzkebski, BrunaMorrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, pages 2790–2799
- Kiana Kheiri and Hamid Karimi, 2023, SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning, arXiv:2307.10234v2 [cs.CL] 23 Jul 2023
- Constantin Orasan. 2003. PALinkA: a highly customizable tool for discourse annotation. In Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, pages 39 - 43, Sapporo, Japan.
- Jonas Pfeiffer, Aishwarya Kamath , Andreas Ruckle , Kyunghyun Ch, Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In Proceedings 365 of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 487–503).
- Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, Guoyin Wang, 2023,Sentiment Analysis through LLM Negotiations, arXiv:2311.01876v1 [cs.CL] 3 Nov 2023
- Kogilavani 0Shanmugavadivel, Sai Haritha Sampath, Pramod Nandhakumar, Prasath Mahalingam, Malliga Subramanian, Prasanna Kumar Kumaresan and, Ruba Priyadharshini, 2022, An analysis of machine learning models for sentiment analysis of Tamil code-mixed data, In: Computer Speech & Language Volume 76
- Sharmista, 2020, Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study, https://mkuniversity.ac.in/research/SYNOPSIS/P4249.pdf
- Thavareesan .S and Mahesan .S, 2019. "Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation," In Proceedings of 14th Conference on Industrial and Information Systems (ICIIS), Kandy, Sri Lanka, 2019, pp 320-325
- Ruan Visser and Marcel Dunaiski, 2022, Sentiment and intent classification of in-text citations using BERT, In Proceedings of 43rd Conference of the South African Institute of Computer Scientists and Information Technologists, EPiC Series in Computing, Volume 85, 2022, Pages 129–145
- Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang1, 2016, Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model, In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 225–230
- Xingyou Wang, Weijie Jiang, Zhiyong Luo3, 2016,Combination of Convolutional and Recurrent Neural Network for Sentiment Analysis of Short Texts, In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2428–2437, Osaka, Japan, December 11-17 2016.
- Yue Zhang and Duy Tin Vo. 2016. Neural Networks for Sentiment Analysis. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, Austin, Texas. Association for Computational Linguistics.

Opportunities and Challenges in Sentiment Analysis and Emotion Recognition in Tamil Text and Speech

Ramesh Kumar V, Krishnan P

ABSTRACT

Sentiment Analysis and Emotion Recognition have become pivotal in understanding human expression and engagement in digital communication. As language enthusiasts and vernacular AI experts keep putting efforts in improving the algorithms, it is essential to explore methodologies and understand the opportunities and challenges in applying some of these techniques to Tamil, fostering inclusive advancements. Tamil exhibits a high degree of morphological complexity with a vast categorization of morphemes semantically. This complexity poses a challenge in accurately deciphering the sentiment expressed as words - as it may change form based on context, requiring sophisticated natural language processing techniques. Polysemy and homonymy are extremely prevalent as one delves deeper into the literature and disambiguating between these meanings is very crucial. Tamil is also spoken in various regions with linguistic variations in dialects and accents. The same gets reflected in the origins of various literary works over different periods of time. So, its imperative that the specific dialect and accent of specific regions based on the literary origins are also considered as a variable in deciphering the exact meaning of the specific literature. Adapting sentiment analysis models to these regional nuances also adds to the complexity and hence interpreting the sentiment of the particular literature.

Most importantly, availability of labeled datasets for training sentiment analysis models in Tamil, even though is made available in the recent times, it is very limited. Creating high-quality, domain-specific labeled datasets is essential for the development and validation of accurate sentiment analysis models. This paper aims to give a complete overview of challenges specific to Tamil sentiment, considering linguistic intricacies, contextual variations, morphological complexity, the richness of Tamil literary expression along with challenges in code-switching. Additionally, the paper also tries to provide a brief overview of emotion recognition in Tamil speech, recognizing the need for a nuanced understanding of emotional cues in spoken Tamil. Analysing some popular speech processing technologies, we explore the development of some of the models attuned to the acoustic characteristics and emotional markers unique to the Tamil language.

Ramesh Kumar V
zSpaze Technologies Private Limited, Kaggadasapura, Bangalore. Email: rkvconf@gmail.com,
Krishnan P
Department of Computer Science and Engineering, Government College of Engineering, Salem
Email: vkrishnan569@gmail.com

I. INTRODUCTION

The 26th edition of Ethnologue reveals that our world is home to approximately 7,168 contemporary living languages. This extensive linguistic diversity highlights the richness with which human communication has evolved over thousands of years, showcasing the myriad ways in which cultures express themselves through unique linguistic structures. The multitude of languages reflects the global mosaic of traditions, histories, and identities, each one contributing to the vibrant spectrum of human expression - varied by the different periods, emotions and a wide array of logic. As we acknowledge this linguistic abundance, it becomes evident that language is not merely a tool for communication but a repository of cultural heritage and a testament to the rich complexity of the human experience.

Major South Indian languages, including Tamil, Telugu, Malayalam, Kannada, and Tulu, exhibit the distinctive characteristic of being agglutinative. In the realm of linguistics, agglutinative languages are characterized by the formation of complex words through the amalgamation of morphemes without any alteration in spelling or phonetic structure.

As one of the world's oldest and classical languages, Tamil boasts a profound literary heritage that spans centuries. With roots deeply embedded in history, Tamil has evolved into a vibrant cultural force, weaving its linguistic tapestry through the narratives of diverse communities extending beyond geographical boundaries, finding expression not only in the hearts of its native speakers but also resonating with a global audience. The agglutinative nature of such South Indian languages, especially Tamil, enhances their flexibility and adaptability, as speakers can effortlessly construct new words by affixing morphemes to convey nuanced meanings. This linguistic structure not only facilitates effective communication but also reflects the rich linguistic diversity and complexity present in South Indian cultures. The preservation of the original spelling and phonetic integrity throughout the word formation process distinguishes agglutinative languages from other linguistic types, contributing to the unique linguistic landscape of the South Indian linguistic milieu.

II. CHALLENGES WITH SYNTAX AND SEMANTICS

Within the realm of linguistics, as like every other language, Tamil also exhibits a dual nature, comprising both form (encapsulated by syntax), and meaning (encompassing semantics and pragmatics). There are several challenges with syntax and semantics even for the most native resident speaking Tamil as his mother tongue.

"Tholkaappiyam", composed by Tholkaappiyar approximately 2700 years ago, stands as an ancient literary masterpiece to better categorise and educate about the syntax, semantics and pragmatics, representing a monumental literary and linguistic asset with enduring importance for Tamil language enthusiasts globally. Although commonly categorized as a grammar guide, its impact transcends the boundaries of linguistic analysis, sentiment analysis and emotion recognition in Tamil text - reaching into broader realms of cultural and literary significance. Given the authenticity and wide acceptance, the challenges of Tamil language in terms of Syntax and semantics as well as sentiment analysis and emotion recognition, takes into account, lot of references from Tholkaappiyam. Let us first look at some of the challenges in terms of syntax and semantics.

A. Polysemy and Homonymy

As one immerses oneself in the vast expanse of Tamil language literature, the prevalence of polysemy and homonymy becomes increasingly apparent, underscoring the intricacies inherent in the language. While there are innumerable examples spread over the sangam literature, below are a couple of examples for the benefit of the readers.

கோ king, cow, sky, heaven, to stitch, pot maker	அடி step, to beat, measure, foot, bottom, antique, lineage

Tamil also exhibits a higher prevalence of homonymy compared to any other language globally. The linguistic landscape of Tamil is uniquely characterized by a multitude of words sharing identical spellings but possessing distinct meanings, contributing to the intricate tapestry of its vocabulary. This prevalence of homonymy adds a layer of complexity to the language, requiring speakers to rely on contextual cues for accurate interpretation. Here are some examples.

(paanam) பானம் - a drink பாணம் - arrow	(palli) பள்ளி - school பல்லி - lizard
(vili / vizhi) விளி - to say விழி - eye	(thalai) தலை - head தளை - a plant base

B. *Rich Morphology*

Tamil showcases an impressive morphological richness, enabling the generation of a multitude of words by incorporating morphological suffixes onto a single root word. This linguistic feature underscores the language's capacity for intricate word formation, where the addition of suffixes to a base word contributes to its expansive vocabulary.

For example,

அவனால்	அவனது	ஆனால்
அவன் + ஆல்	அவன் + அது	ஆ + நால்
avan + aal	avan + athu	aa + naal
through him.	it's his.	but

C. Lexical Diversity

Tamil language also stands out for its remarkable lexical diversity, presenting an abundance of words that convey similar meanings. The extensive lexical diversity in Tamil enhances its adaptability and precision, allowing for subtle variations in meaning that cater to the diverse contexts in which the language is employed. The word "கோ" as mentioned in section II.A is one of the classical examples that can have several meanings and synonymous expressions depending on the context of usage.

D. Syntactic / Lexical Ambiguity

The syntax of Tamil language plays a pivotal role in conveying meaning, ensuring clarity, and fostering coherent discourse. As like other natural languages, Tamil also has its share of syntactic ambiguity.

The fundamental structure of sentences in Tamil primarily adheres to a Subject-Object-Verb format, but can also be conveyed in a Object-verb format without the need for Subject, especially with statements denoting gender. In this linguistic framework, the inclusion of an subject is not mandatory, offering flexibility in expression. For example, in a

[1] அறை கொடுத்தான். (or)

[2] அவன் அறை கொடுத்தான்.

The ambiguity in both the sentences [1] & [2] arises from the several possible meanings of the word "அறை" which can refer to either "a half" (or) "slap" (or) "room" (or) "answer" (specifically in a literary sense) depending on the context. This ambiguity

becomes more pronounced in larger structures where constituents may have multiple interpretations based on their internal structure and syntactic position.

In addition, both [1] and [2] denote the same meaning that "he gave" something (depending on the context in which the statement is used. In the first sentence, even though the Subject "அவன்" is absent, it still conveys the same meaning. This is one of the peculiar challenges in Tamil literature. Consider another example of lexical ambiguity, where a word holds multiple meanings.

கருப்பு மருந்து குப்பி

The ambiguity in the sentence stems from the term "கருப்பு" which can attribute itself to either "மருந்து" meaning "medicine" or "குப்பி" meaning "bottle." The first instance pertains to homonyms or homographs, contributing to lexical ambiguity, while the second exemplifies structural ambiguity, where the sentence's structure allows for multiple interpretations. Lexical ambiguity, induced by homonyms or homographs, underscores the challenge of precise interpretation in language, whereas structural ambiguity adds complexity by permitting various meanings based on syntactic arrangements.

III. CHALLENGES WITH SENTIMENTS & EMOTIONS

Starting from sanga time period till today, Most of the eminent Tamil scholars and linguists have an implicit knowledge or competence to understand the ambiguous utterances of syntax and semantics. In the same manner, mostly the Tamil scholars and linguists clearly understand the intricacies of emotional expressions throughout the Tamil literature. But for the understanding of normal people, Tholkappiam details the eight different types of emotions.

E. Vast pool of expressed emotions

One of the key challenges of emotion recognition is the identification and training of all the human emotions. This section of the paper once again refers to Tholkappiam and delves into the concept of "Meipadugal" which represents a total of 32 emotions throughout Tamil literature. These are then narrowed down to eight primary or distinct emotions intricately woven throughout Tamil literary expressions. These emotions are listed in a poem in Tholkappiyam.

நகையே அழுகை இளிவரல் மருட்கை
அச்சம் பெருமிதம் வெகுளி உவகை யென்று
அப்பால் எட்டே மெய்ப்பாடு என்ப. (தொல். மெய்ப். 3)

The overall list of eight primary or distinct human emotions in the larger context can be summarised as below.

1. நகை, nagai (joy),
2. அழுகை, azhugai (sadness),
3. இளிவரல், ilivaral (disgust),
4. மருட்கை, marutkai (confusion),
5. அச்சம், accham (fear),
6. பெருமிதம், perumidhan (pride)
7. வெகுளி, vekuli (hate)
8. உவகை, uvakai (love)

These eight Meipadugal encapsulates a profound summary of cultural and linguistic understanding, serving as a vital anchor for the analysis of sentiment and emotion recognition specific to the Tamil language.

F. Challenge with Sarcasm identification

Satire or Sarcasm is a composition that provides a deep rooted meaning. In Tamil Literature, this term has undergone various interpretations by different renowned poets and writers. It has been employed as a form of prayer coupled with insults (Nindastuthi or நிந்தாஸ்துதி) and, at other times, utilized as wordplay and clever phrases (Sladai or சிலேடை). On certain occasions, eminent poets exercised caution to conceal their mockery through phrasing that did not overtly reveal their intent. The term "Satire" has evolved to represent a poignant commentary, akin to a bitter pill enveloped in the sweetness of language or, conversely, a sugar-coated critique that conceals its acerbic nature. Sarcasm, as a linguistic expression, frequently defies conventional language norms by communicating a meaning totally contrary to its literal interpretation. The nuanced nature of sarcasm presents a challenge in deciphering its intricacies, particularly in multilingual environments where diverse cultural nuances come into play. The ability to recognize sarcasm is paramount for ensuring precise sentiment analysis and comprehending the subtle nuances embedded in communication.

In the realm of AI based language processing, sarcasm introduces complexity as it relies heavily on context, tone, and cultural subtleties. In certain contexts, it conveys layered meanings, often requiring a deeper understanding of the speaker's intent, the time of publishing and the geographical significance of the publication. The intricacies of sarcasm highlight the dynamic nature of language and the need for sophisticated language models that can adeptly navigate the complexities presented by sarcastic expressions. In multilingual settings, where cultural contexts vary, recognizing and interpreting sarcasm becomes even more intricate, underscoring the importance of linguistic models capable of discerning the nuanced layers of communication for accurate sentiment analysis.

For example, a well renowned poet in one of his sarcastic poems about Lord Shiva, had written as below.

நச்சரவம் பூண்டதில்லை நாதரே
தேவரீர்பிச்சையெடுத் துண்ணப் புறப்பட்டும் – உச்சதமாங்
காளனேன் குஞ்சரமேன் கார்க்கடல்போற் றான்முழங்கும்
மேளமேன் ராசாங்க மேன். (கா. 65)

The actual meaning is different but the hidden சிலேடை meaning is different.

The poem goes on like praising Lord Shiva on his glamorous procession around the streets but it actually talks about the procession actually for begging. Analysis of such literary works requires advanced AI models.

G. Challenges with Dialects, Accents & Period works

There is also a considerable amount of linguistic variations manifested in distinct dialects and accents. This linguistic diversity is intricately intertwined with the origins of literary works spanning different epochs, each imbued with unique regional characteristics. Consequently, it becomes imperative to recognize and incorporate the specific dialects and accents associated with distinct literary origins and the time frame - as variables when deciphering the precise meaning of literature.

For example, In the Sangam period, the term “Veguli” (வெகுளி) held a distinct meaning, referring to anger or hate as depicted in the couplet

“வெகுளி கணமேனுங் காத்தலரிது” (Kural, 29).

During that particular era, there existed a verb form, “Vegul” or “Vegulthal” (வெகுள்/வெகுள்தல்), akin to “Nagai/Nagaithal,” (நகை/நகைத்தல்) actively used. While the verb form has ceased to be in use today, the noun form “Veguli” has undergone a semantic shift. It no longer signifies anger but has adopted an entirely new meaning—’innocent or naive.’

The vestiges of the original verb “Vegul” linger in words such as “Vegundan” (வெகுண்டான்), denoting irritation. Notably, only the past tense is prevalent, lacking present or future tense conjugations like “Vegulkiran” (வெகுள்கிறான்) or “Vegulvan” (வெகுள்வான்), though grammatically plausible.

The absence of these forms stems from the linguistic evolution wherein the original verb “Vegul” has become obsolete, and the noun “Veguli” has assumed a new connotation. Presently, “Vegundan” is employed as a figure of speech rather than a consciously used verb conjugation, reflecting the linguistic transformations and shifts in meaning that have occurred over time.

Another popular term that is commonly used in tamil movie songs is “Annam” (அன்னம்/அன்னப்பறவை). In contemporary usage, “Annam” is commonly associated with swans. However, this was not the case during the Sangam period. Numerous descriptions in Sangam literature indicate that “Annam” actually refers to the common teal, a lesser bird-of-paradise, or a kind of a whistling duck.

Another instance of a word losing its original meaning due to misinterpretation is “Kavarima” (கவரிமா). Over time, this term was incorrectly perceived as “Kavarimaan,” (கவரிமான்) leading to its classification as a mythological creature. In reality, “Kavarima” refers to an almost extant species called Yak (which is mostly found in the himalayan regions), emphasizing the impact of misinterpretations on linguistic understanding.

Similarly, “Kozhunan” (கொழுநன்) originally denoted a husband, as evidenced in the usage found in Silapathigaram(சிலப்பதிகாரம்). However, in contemporary language, “Kozhunan” is often mistaken for brother-in-law, and it is plausible that “Kozhunthan” (கொழுந்தன்) may have derived from “Kozhunan,” showcasing how linguistic nuances and meanings can shift over time due to misinterpretations and evolving usage.

Such regional and period specific nuances, dialects and accents pose a significant challenge in the realm of developing AI models for sentiment analysis.

IV. OVERVIEW OF TECHNIQUES FOR EMOTION & SENTIMENT ANALYSIS

Over the years, progress in computing has significantly contributed to the evolution and refinement of Tamil language processing. The trajectory of Tamil computing, spanning from shallow parsing to machine translation, has witnessed notable advancements. Numerous initiatives have been undertaken to create technological tools and applications specifically tailored for the Tamil language (Rajendran, S., et.al, 2018). Educational institutions, independent researchers, Tamil enthusiasts, and both national and international companies have invested extensive efforts over the years to advance technology dedicated to the Tamil language.

Some of the techniques developed or adapted since last couple of decades include but not limited to:

H. Shallow Parsing

Shallow parsing is an initial phase in language processing that involves the identification of tokens, parts-of-speech, and the delineation of phrasal units or “chunks” within sentences. In the domain of Tamil computing, early endeavours were dedicated to formulating shallow parsing algorithms aimed at

extracting meaningful linguistic units like noun phrases and verb phrases from Tamil text. Researchers utilized rule-based methodologies and linguistic heuristics, laying the foundational groundwork for more sophisticated language processing techniques.

I. Tokenization

Tokenization, a crucial process, entails breaking a text into smaller units referred to as tokens. These tokens can encompass fully formed words, sub words, characters, or phrases, contingent upon the desired granularity of analysis. While spaces play a pivotal role in token identification in written language, Tamil presents a challenge with frequently occurring multi-token words (MTW) that require segmentation for subsequent processing steps. MTWs introduce multiple grammatical components within a single word, necessitating a token status for further analysis.

J. Part-of-Speech (POS) Tagging

Part-of-speech (POS) tagging, another fundamental aspect, assigns grammatical categories such as nouns, verbs, and adjectives to words, facilitating deeper analysis and comprehension of sentence structure. These tagged elements serve as vital building blocks for subsequent natural language processing tasks. It involves open class category tagging like adjective, adverb, pronoun, verb, interjection, noun, etc., and closed class category tagging like coordinating conjunction, adposition, numeral, auxiliary, determiner, particle, subordinating conjunction, proper noun etc.

K. Chunking

Chunking involves grouping smaller units or constituents in a sentence to comprehend their phrasal structure, aiding in the identification of phrases and relationships and offering insights into syntax and language comprehension.

L. Morphological Analysis

Morphological analysis stands as another integral component in Tamil language processing, encompassing the breakdown of words into morphemes. This process aids in understanding word structures and inflections in Tamil, where inflections, representing modifications or additions to a word, convey crucial grammatical and contextual information. For instance, Tamil employs inflections for verb conjugations, tense markers, case markers, and gender agreements. Analyzing these inflections unveils the syntactic and semantic nuances embedded in the language, fostering more accurate and insightful language processing. The incorporation of universal dependency morph features becomes pivotal, as these features, realized as inflections within words in Tamil, play a crucial role in language understanding and interpretation.

V. OVERVIEW OF PAST WORKS (INDICATIVE)

Sobha Lalitha Devi et al. implemented a finite state automata model to capture the regular inflectional patterns in Tamil, enhancing the accuracy of morpheme extraction. The processing of a Tamil word unfolds from right to left, with the finite state automaton accepting suffixes and progressing until the final state reveals the root word. A valid word is identified when morphemes are accepted at all states, triggering the extraction of its morphological analysis.

Mokanarangan et al. devised a morphological engine designed to generate multiple candidate analyses for a word, leveraging an annotated lexicon corpus and a repository of Tamil grammatical rules. These candidates underwent assessment by an SVM classifier, utilizing a high-frequency word list to predict the most accurate analysis. The SVM's considerations encompassed frequency-based scores, suffixes, lexical labels, and average length as features, contributing to a robust morphological analysis system.

A guidebook titled "Building Transformer Models with Attention," authored by Jason Brownlee, Stefania Cristina, and Mehreen Saeed in 2022 for Machine Learning Mastery, serves as a comprehensive resource. The authors illustrate the process of creating a Neural Machine Translator (NMT) from the ground up using the transformer architecture in Keras. This guide offers practical insights and implementation techniques, particularly emphasizing attention mechanisms for natural language processing in complex languages like Tamil. Tailored for developers and machine learning enthusiasts, this resource provides hands-on guidance for constructing advanced models, showcasing real-world applications of transformer architectures, particularly in the realm of machine translation.

Nazir and Wang conducted an extensive survey, delving into the dynamic landscape of ChatGPT. Their focus encompassed the advancements, applications, prospects, and challenges associated with a diverse array of linguistic computations.

VI. BRIEF OVERVIEW OF SENTIMENT ANALYSIS ALGORITHMS

While specific sentiment analysis algorithms tailored exclusively for the Tamil language may not be as abundant as those for more widely spoken languages, the following list includes general sentiment analysis algorithms and approaches that can be adapted or trained for Tamil sentiment analysis:

VADER (Valence Aware Dictionary and sEntiment Reasoner) - A rule-based sentiment analysis tool that

is particularly effective for social media text. It can be customized for sentiment analysis in Tamil.

TextBlob - Offers a simple API for diving into common natural language processing (NLP) tasks, including sentiment analysis. It can be trained or adapted for Tamil sentiment analysis.

FastText - Developed by Facebook, FastText is a library for efficient learning of word representations and sentence classification. It can be trained for sentiment analysis in Tamil.

BERT (Bidirectional Encoder Representations from Transformers) - While BERT models are often pretrained on large datasets in English, fine-tuning on smaller datasets for Tamil sentiment analysis can be explored.

LSTM (Long Short-Term Memory) - A type of recurrent neural network (RNN) that can be used for sequence prediction problems, including sentiment analysis. It requires labelled datasets for training in Tamil.

Word2Vec - Represents words in vector space, which can capture semantic relationships. Pretrained models can be used or adapted for Tamil sentiment analysis.

Naive Bayes - A probabilistic algorithm that makes assumptions about the independence of words. It can be trained for sentiment analysis in Tamil.

SVM (Support Vector Machines) - A machine learning algorithm that can be applied to sentiment analysis tasks. It requires labeled training data in Tamil.

CNN (Convolutional Neural Network) - Effective for text classification tasks, including sentiment analysis. It can be adapted for Tamil sentiment analysis with proper training data.

Gated Recurrent Unit (GRU) - A type of recurrent neural network similar to LSTM, suitable for sequence-based tasks like sentiment analysis in Tamil.

mBERT (Multilingual BERT) is a language representation model developed by Google, pretrained on diverse languages.

MuRIL (Multilingual Representations for Indian Languages) extends mBERT's capabilities to Indian languages, promoting cross-lingual understanding. Both models enhance natural language processing tasks by capturing multilingual nuances and improving performance across various languages.

VII. CONCLUSION

It's important to note that building effective sentiment analysis models for Tamil may require extensive labelled datasets in Tamil. Tremendous amount of efforts and work is being carried out by public and private organisations to increase the dataset pool for classical as well as modern Tamil literature. In addition, the existing algorithms may need fine-tuning or training specifically for the language nuances. Additionally, leveraging transfer learning or pretraining on larger datasets in English followed by fine-tuning on smaller Tamil datasets can be an effective strategy to build efficient language models specifically for Tamil.

REFERENCES

[1] Dhanalakshmi V, Anandkumar M, Rekha RU, Arunkumar C, Soman KP, Rajendran S (2009) Morphological analyzer for agglutinative languages using machine learning approaches. In: Proceedings of international conference on advances in recent technologies in communication and computing, pp 433–435

[2] The Unicode consortium [Online]. Available: https://unicode.org/cldr/charts/latest/supplemen tal/languages_and_scripts.html

[3] Devi SL, Marimuthu K, Ram RVS, Bakiyavathi T, Amudha K (2013) Morpheme extraction in Morpheme extraction in Tamil using finite state machines. Presented at forum for information retrieval evaluation, New Delhi, India, December 4–6, 2013

[4] Bhardwaz, S., & Kumar, J. (2023). An extensive comparative analysis of chatbot technologies - ChatGPT, Google BARD and Microsoft Bing. 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC). https://doi.org/10.1109/icaaic56838.2023.10140214

[5] University College Dublin [Online]. Available: http://www.ucd.ie/artspgs/introling/Aronoffmorphology.pdf

[6] Anand Kumar M, Dhanalakshmi V, Soman KP, Rajendran S (2010) A sequence labeling approach to morphological analyzer for Tamil language. Int J Comput Sci Eng 2(6):1944–1951

[7] Mokanarangan TP, Megala U, Nilusija N, Dias G, Jayasena S, Ranathunga S (2016) Tamil morphological analyzer using support vector machines. In: Proceedings of international conference on applications of natural language to information systems, Salford, UK, pp 15–23

[8] Tunstall, L., Werra, L. V., & Wolf, T. (2022). Natural language processing with transformers. O'Reilly Media.

[9] Building Transformer Models with Attention: Implementing a Neural Machine Translator from Scratch in Keras. Jason Brownlee, Stefania Cristina, Mehreen Saeed, Machine Learning Mastery, 2022.

[10] Maanikkavasaasakan (2010) Tholkaappiyam. Uma Padhippaagam, Chennai.

[11] Clark, A., Fox, C. and Lappin, S. eds., 2012. The handbook of computational linguistics and

natural language processing (Vol. 118). John Wiley & Sons.

[12] De Marneffe, M.C., Manning, C.D., Nivre, J. and Zeman, D., 2021. Universal dependencies. Computational linguistics, 47(2), pp.255-308

[13] Sakirin, T., & Ben Said, R. (2023). User preferences for chatgpt-powered conversational interfaces versus traditional methods. Mesopotamian Journal of Computer Science, 24-31. https://doi.org/10.58496/mjcsc/2023/006

[14] R Kiran, K Nivedha, T Subha, et al. 2017. "Voice and speech recognition in tamil language". In 2017 2nd International Conference on Computing and Communications Technologies (ICCCT), pages 288--292. IEEE.

[15] Ramasamy, L., and Žabokrtský, Z. 2011. Tamil Dependency Parsing: Results Using RuleBased And Corpus Based Approaches. In International Conference on Intelligent Text Processing and Computational Linguistics. Berlin: Springer, pp. 82-95.

[16] Nazir, A., & Wang, Z. (2023). A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. Meta-Radiology, 100022. https://doi.org/10.1016/j.metrad.2023.100022.

[17] Renganathan, Vasu 2016. Computational Approaches to Tamil Linguistics. Cre-A, Chennai.

[18] தொல்காப்பியர் எனும் மெய்ப்பாட்டியல் அறிஞர்! By Munaivar Ko.Vijayaraghavan | Published in Dinamani. Published On : 02nd July 2023, https://www.dinamani.com/weekly-supplements/tamilmani/2023/jul/02/tolkappiyar-is-a-true-scholar-4031133.html

[19] உடையார், கவி காளமேகம் – சிலேடை பாடல்கள், September 8, 2011 in தமிழும் நயமும்

[20] Mark Purdy, John Zealley, and Omaro Maseli, 2019, The Risks of Using AI to Interpret Human Emotions, Published in Harvard Business Review

[21] Chevillard, Jean-Luc, 2010b, " 'Rare words' in classical Tamil literature: from the Uriyiyal to the Tivākaram ", pp. 301-317, in ACTA ORIENTALIA, Volume 63, Number 3/September 2010

[22] S.Rajendran, Resolution of Lexical Ambiguity in Tamil, published in the Computational linguistics and NLP Amrita Vishwa Vidya Peetam, Coimbatore.

[23] Sarveswaran, K., Dias, G. & Butt, M. ThamizhiMorph: A morphological parser for the Tamil language. Machine Translation 35, 37–70 (2021). https://doi.org/10.1007/s10590-021-09261-5

[24] Agesthialingom, S. (1971). A note on Tamil verbs. Anthropol Linguist, 13.

[25] Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., & Mohrir, M. (2007). OpenFst: a general and efficient weighted finite-state transducer library BT - International conference on implementation and application of automata. Springer. https://doi.org/10.1007/978-3-540-76336-9_3

[26] M., Suriyah., Aarthy, Anandan., Anitha, Narasimhan., Madhan, Karky. (2019). Piripori: Morphological Analyser for Tamil. doi: 10.1007/978-3-030-24051-6_75

[27] S., Lushanthan., A., R., Weerasinghe., Dulip, Herath. (2014). Morphological analyzer and generator for Tamil Language. 190-196. doi: 10.1109/ICTER.2014.7083900

[28] Rajasekar, M., Geetha, A. (2022). Comparison of Machine Learning Methods for Tamil Morphological Analyzer. In: Raj, J.S., Palanisamy, R., Perikos, I., Shi, Y. (eds) Intelligent Sustainable Systems. Lecture Notes in Networks and Systems, vol 213. Springer, Singapore. https://doi.org/10.1007/978-981-16-2422-3_31

[29] S., Lushanthan., A., R., Weerasinghe., Dulip, Herath. (2014). Morphological analyzer and generator for Tamil Language. doi: 10.1109/ICTER.2014.7083900

[30] Ananthi, Sheshasaayee., V., R., Angela, Deepa. (2016). A Conceptual Model for Acquisition of Morphological Features of Highly Agglutinative Tamil Language Using Unsupervised Approach. doi: 10.1007/978-81-322-2757-1_49

Sentiment Analysis and Emotion Recognition in Tamil Text and Speech using a SVM-RF Integrated Model

Hareni Srikanth, Harshadha K S, Hajeera Thabasum A, Rajkumar Kalaimani

ABSTRACT

Abstract for sentiment analysis and emotion recognition within the context of Tamil language, encompassing both text and speech data. Leveraging natural language processing techniques, the research aims to develop robust models capable of discerning sentiments and identifying emotions expressed in Tamil content. The investigation incorporates diverse datasets, including written text and spoken language, to enhance the model's adaptability. Methodologies involve feature extraction, machine learning algorithms, and deep learning architectures to capture nuanced emotional nuances unique to Tamil communication. The outcomes are anticipated to contribute to advancements in sentiment analysis and emotion recognition, fostering applications in diverse domains such as customer feedback analysis, social media monitoring, and human-computer interaction tailored to Tamil speakers.

Hareni Srikanth
Adhiyamaan College of Engineering, Hosur.
harenisrikanth23@gmail.com

Harshadha K S
Adhiyamaan College of Engineering. Hosur.
harshadhadestiny@gmail.com

Hajeera Thabasum A
Adhiyamaan College of Engineering, Hosur.
hajeerathabasum281@gmail.com

Rajkumar Kalaimani (Mentor),
Senior Engineer - Product and Platform Engineering,
Altimetrik India Pvt Ltd, Chennai.
akr.rajkumar@gmail.com

INTRODUCTION

The field of sentiment analysis and emotion recognition has witnessed substantial growth in recent years, driven by the increasing demand for understanding human sentiments in diverse linguistic contexts. While numerous studies have explored these domains in widely spoken languages, there exists a significant gap when it comes to languages with distinct linguistic characteristics, such as Tamil. This research endeavours to bridge this gap by delving into the intricate nuances of sentiment and emotion within the context of the Tamil language, encompassing both written text and spoken language.

Tamil, a Dravidian language spoken predominantly in the Indian subcontinent, boasts a rich literary tradition and cultural diversity. Its unique syntactical structures and linguistic idiosyncrasies present a compelling challenge for sentiment analysis and emotion recognition systems. Recognizing the cultural and contextual aspects embedded in the Tamil language is crucial for developing models that can accurately capture the sentiments and emotions expressed by Tamil speakers.

In the era of natural language processing (NLP), the significance of understanding sentiments goes beyond mere linguistic analysis. It extends to applications in customer feedback analysis, social media monitoring, and human-computer interaction tailored to specific linguistic communities. Therefore, the primary objective of this research is to leverage advanced NLP techniques to develop robust models capable of discerning sentiments and identifying emotions in Tamil content.

The investigation adopts a comprehensive approach, incorporating diverse datasets that mirror the linguistic richness of Tamil communication. The datasets encompass both written text, sourced from social media, news articles, and online forums, and spoken language, collected through interviews and recordings. This dual-modal dataset ensures that the developed models are adaptive to the varied forms of expression within the Tamil language, be it the written word or the nuances embedded in spoken discourse.

Methodologically, this research employs advanced feature extraction techniques to capture the distinctive characteristics of Tamil sentiments and emotions. For

written text, we employ TF-IDF and word embeddings, while for spoken language, we leverage Mel-Frequency Cepstral Coefficients (MFCCs) to represent the acoustic features of speech. The choice of these techniques is driven by the need to preserve the cultural and linguistic richness of Tamil expressions.

In the realm of model development, we explore both traditional machine learning algorithms and state-of-the-art deep learning architectures. Support Vector Machines (SVM) and Random Forests are considered for their interpretability and effectiveness in handling textual data. Concurrently, deep learning models, including Recurrent Neural Networks (RNNs) and Transformer architectures, are employed to capture sequential dependencies and semantic relationships within the Tamil language.

The anticipated outcomes of this research extend beyond the academic sphere. We envision that the developed models will find practical applications in domains such as customer feedback analysis, where understanding customer sentiments is paramount for business growth, and social media monitoring, where real-time sentiment analysis aids in gauging public opinions. Moreover, the incorporation of emotion recognition in human-computer interaction tailored to Tamil speakers has the potential to enhance user experiences in technology-driven applications.

ARCHITECTURE

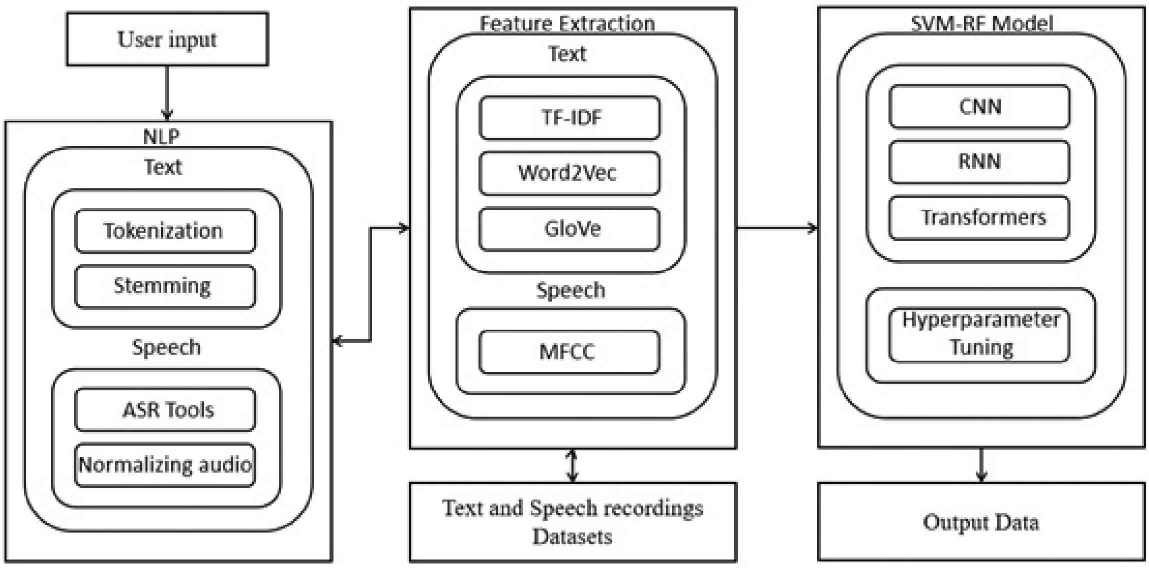


Figure 1: Architecture of the model

The architecture of our sentiment analysis and emotion recognition model is designed to capture the intricate nuances of the Tamil language, employing a dual-modal approach that incorporates both written text and spoken language. The model is divided into key components, beginning with data preprocessing and feature extraction, followed by the implementation of traditional machine learning and state-of-the-art deep learning algorithms.

DATA PREPROCESSING AND FEATURE EXTRACTION

The initial phase involves the preprocessing of both textual and speech data. For written text, we employ techniques such as tokenization, stemming, and the removal of stop words to clean and standardize the input. Additionally, we utilize TF-IDF and word embeddings, such as Word2Vec and GloVe, to extract relevant features that capture the semantic relationships and contextual meaning embedded in the Tamil language. In the case of spoken language, the audio data undergoes preprocessing through the extraction of Mel-Frequency Cepstral Coefficients (MFCCs), capturing the acoustic features necessary for discerning emotional nuances.

MACHINE LEARNING COMPONENT

The model integrates traditional machine learning algorithms to analyze the preprocessed textual features. Support Vector Machines (SVM) and Random Forests are chosen for their ability to handle high-dimensional data and their interpretability. These algorithms are trained on the extracted features, learning to map the linguistic and emotional patterns present in the Tamil language.

DEEP LEARNING COMPONENT

Simultaneously, the model incorporates deep learning architectures to capture the sequential dependencies within the data. Recurrent Neural Networks (RNNs) and Transformer models are employed to process both textual and speech data. RNNs are adept at capturing temporal dependencies, while Transformer models excel in handling non-sequential relationships within the data. This dual approach ensures that the model is capable of recognizing complex emotional patterns present in both written and spoken Tamil expressions.

MODEL FUSION AND OUTPUT

The outputs from the machine learning and deep learning components are fused at an integration layer. This fusion is essential for synthesizing the diverse insights gained from the two modalities, creating a comprehensive understanding of sentiments and emotions in Tamil content. The final output provides not only sentiment labels but also nuanced emotion predictions, offering a more holistic interpretation of the emotional landscape within the Tamil language.

TRAINING AND OPTIMIZATION

The entire model undergoes a rigorous training phase using diverse datasets, and hyperparameters are fine-tuned to enhance performance. The training process involves minimizing the loss function by adjusting model parameters, ensuring that the model generalizes well to unseen data. Optimization techniques such as dropout and batch normalization are applied to prevent overfitting and improve the model's robustness. The deployed model is integrated into real-world applications, allowing end-users to benefit from sentiment analysis and emotion recognition tailored to the Tamil language. Continuous monitoring and feedback mechanisms are implemented to gather user insights, facilitating future improvements to the model. The modular and adaptable architecture ensures that the model can be extended to accommodate additional linguistic complexities and evolving language patterns in Tamil communication.

METHODOLOGY

Data Collection

The foundation of our methodology lies in the collection of diverse datasets that accurately represent the linguistic and emotional landscape of the Tamil language. For written text, we employ web scraping techniques to gather data from social media platforms, news articles, and online forums. Simultaneously, spoken language data is collected through interviews and recordings, capturing the varied expressions and tones inherent in natural speech. This dual-modal dataset ensures a comprehensive understanding of sentiment and emotion in both written and spoken Tamil communication.

Data Preprocessing

Before feeding the data into the model, an extensive preprocessing phase is implemented. For written text, we employ tokenization, stemming, and the removal of stop words to clean and standardize the textual content. In the case of spoken language, the audio data undergoes preprocessing, including the conversion of speech to text using Automatic Speech Recognition (ASR) and normalization of audio signals. This ensures that the input data is ready for subsequent feature extraction and analysis.

Feature Extraction

Feature extraction is a critical component of our methodology, where we aim to capture the linguistic and acoustic characteristics that define sentiment and emotion in Tamil. For written text, we utilize techniques such as TF-IDF and word embeddings (Word2Vec, GloVe) to represent the semantic relationships between words. In the realm of spoken language, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted to represent the acoustic features crucial for discerning emotional nuances in speech.

Model Development

The model development phase incorporates both traditional machine learning algorithms and deep learning architectures. For traditional machine learning, Support Vector Machines (SVM) and Random Forests are chosen for their interpretability and effectiveness in handling textual data. Simultaneously, deep learning models, including Recurrent Neural Networks (RNNs) and Transformer architectures, are employed to capture sequential dependencies within the data. This dual-model approach ensures a comprehensive understanding of sentiment and emotion across diverse linguistic expressions.

Model Training and Evaluation

The models are trained on the prepared datasets, with rigorous evaluation using techniques such as cross-validation to ensure robustness and generalizability. During training, the models learn to map the extracted features to sentiment labels and emotion predictions. Hyperparameter tuning is performed to optimize model performance, and the models are evaluated on metrics such as accuracy, precision, recall, and F1 score to gauge their effectiveness in capturing the nuances of Tamil sentiment and emotion.

Integration and Deployment

Once the models demonstrate satisfactory performance, they are integrated into real-world

applications. Cloud services such as AWS or Azure are employed for scalable and efficient deployment. Web development frameworks are used to create user interfaces that allow seamless interaction with the models, facilitating practical applications in domains such as customer feedback analysis and social media monitoring.

Testing and Optimization

The deployed models undergo thorough testing to ensure their reliability and accuracy in real-world scenarios. Quality assurance tools are employed, and continuous monitoring mechanisms are established to gather user feedback. Model optimization techniques, including pruning and quantization, may be applied to enhance efficiency and reduce computational resources while maintaining accuracy.

SVM-RF Model

Support Vector Machines (SVM)

The Support Vector Machines (SVM) component of our hybrid model is a classical machine learning algorithm renowned for its effectiveness in binary and multiclass classification tasks. In the context of sentiment analysis and emotion recognition for Tamil language, SVM serves as a robust foundation. SVM works by finding the optimal hyperplane that separates data points of different classes with a maximum margin. In our model, the SVM classifier is trained on the preprocessed textual features extracted from the written Tamil content. By mapping these features onto a higher-dimensional space, SVM strives to discern the complex patterns and boundaries inherent in sentiment-laden textual data.

Random Forests (RF)

Complementing the SVM component is the Random Forests (RF) algorithm, a powerful ensemble learning method that excels in handling complex, high-dimensional data. Random Forests operate by constructing a multitude of decision trees during the training phase, each tree offering its prediction. In the case of our model, the RF component is trained on the same textual features as SVM, providing a diversified perspective on the learned patterns. By aggregating the predictions of multiple decision trees, Random Forests mitigate overfitting and enhance the model's generalization capabilities. This is particularly valuable for capturing the diverse linguistic expressions present in Tamil text.

Integration and Fusion

The SVM and RF components operate in parallel, each providing its unique insights into the sentiment and emotion expressed in Tamil content. The outputs from these classifiers are then fused at an integration layer, allowing for a synergistic interpretation of the dual-modal data. This fusion is crucial for creating a holistic understanding of sentiments and emotions, as SVM and RF capture different aspects and subtleties in the complex landscape of the Tamil language. By combining the strengths of both algorithms, our SVM-RF model strives to achieve a more comprehensive and nuanced analysis than what could be attained with either algorithm in isolation.

Hyperparameter Tuning

Both SVM and RF models undergo meticulous hyperparameter tuning to optimize their performance. Parameters such as the regularization parameter (C) in SVM and the number of trees and maximum depth in RF are fine-tuned using techniques like grid search or random search. This ensures that the models generalize well to unseen data and are capable of capturing the specific linguistic nuances present in Tamil sentiments and emotions.

Application in Sentiment Analysis and Emotion Recognition

The SVM-RF model is applied to real-world applications such as customer feedback analysis and social media monitoring, leveraging its ability to discern sentiments and emotions expressed in both written text and spoken language. The model's output, synthesized through the fusion of SVM and RF predictions, provides a more comprehensive understanding of the emotional landscape within the Tamil language. The adaptability of this hybrid model makes it well-suited for applications where capturing the richness and diversity of emotions in Tamil communication is paramount.

PROCESS FLOW

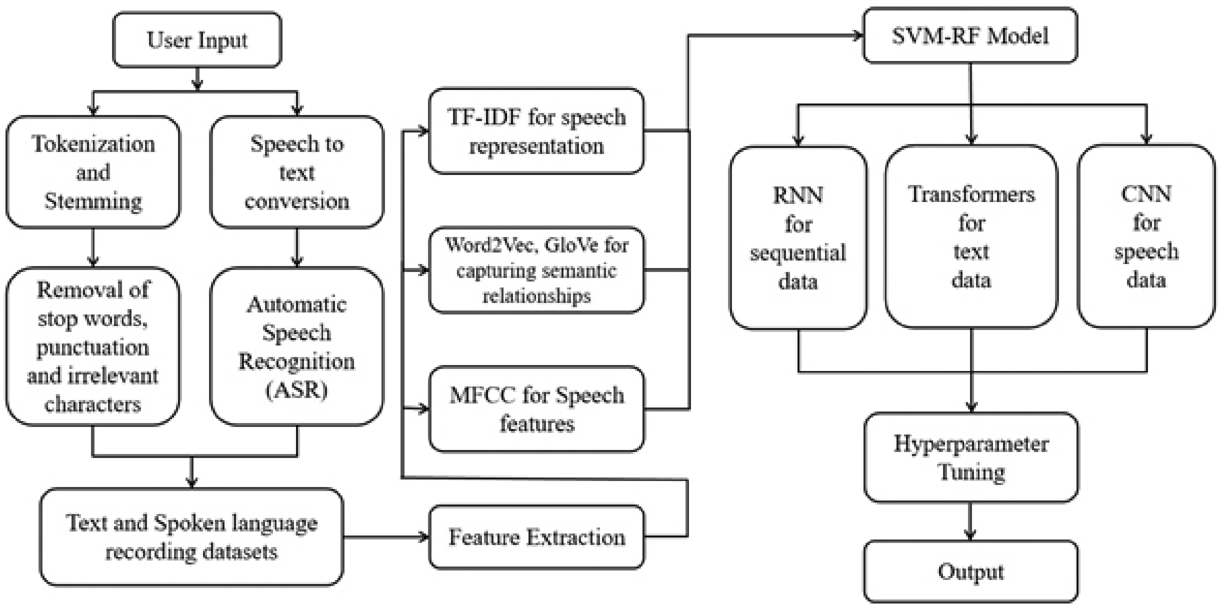


Figure 2: Process flow of the model

Data Collection

The process begins with the collection of diverse datasets, encompassing both written text and spoken language, to represent the linguistic and emotional richness of the Tamil language. Written text is obtained through web scraping from social media, news articles, and online forums, while spoken language data is collected through interviews and recordings. This dual-modal dataset forms the foundation for training the hybrid SVM-RF model.

Data Preprocessing

The collected data undergoes a rigorous preprocessing phase to ensure that it is ready for feature extraction. For written text, this involves tokenization, stemming, and the removal of stop words to clean and standardize the textual content. In the case of spoken language, audio data is converted to text using Automatic Speech Recognition (ASR), and signals are normalized. This preprocessing step sets the stage for extracting meaningful features that capture the linguistic and acoustic characteristics specific to Tamil sentiments and emotions.

Feature Extraction

Feature extraction is a critical step in our process flow. For written text, we utilize techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings (Word2Vec, GloVe) to represent the semantic relationships and contextual meaning embedded in the Tamil language. Simultaneously, Mel-Frequency Cepstral Coefficients (MFCCs) are extracted from the spoken language data to capture the acoustic features essential for discerning emotional nuances in speech.

SVM Model Training

The preprocessed and feature-extracted data is then fed into the Support Vector Machines (SVM) component of the model. SVM is trained to discern sentiment and emotion patterns within the written Tamil content. The algorithm aims to find the optimal hyperplane that separates different classes of sentiments, learning the complex patterns and boundaries present in the high-dimensional feature space.

RF Model Training

Concurrently, the Random Forests (RF) component is trained on the same textual features as SVM.

Multiple decision trees are constructed during the training phase, each offering its unique prediction. The ensemble of decision trees in the Random Forests model mitigates overfitting and enhances the model's ability to generalize well to diverse linguistic expressions.

Integration of SVM and RF

The outputs from the SVM and RF components are integrated at a fusion layer. This integration allows for the synthesis of the unique insights provided by each algorithm. The combination of SVM and RF

predictions ensures a more comprehensive and nuanced understanding of sentiments and emotions in Tamil content, considering the diverse linguistic expressions present in both written and spoken language.

Hyperparameter Tuning

Both the SVM and RF models undergo hyperparameter tuning to optimize their performance. Parameters such as the regularization parameter (C) in SVM and the number of trees and maximum depth in RF are fine-tuned using techniques like grid search or random search. This step ensures that the models generalize well and effectively capture the specific linguistic nuances of Tamil sentiments and emotions.

Model Deployment

Once the SVM-RF model is trained and fine-tuned, it is deployed for practical applications. Cloud services such as AWS or Azure are employed for scalable and efficient deployment. Web development frameworks are utilized to create user interfaces, allowing end-users to interact with the model and receive sentiment and emotion predictions tailored to the Tamil l anguage.

Evaluation and Feedback Mechanism

The deployed model is rigorously evaluated on metrics such as accuracy, precision, recall, and F1 score to assess its performance. Continuous monitoring and feedback mechanisms are established to gather user insights, facilitating future improvements to the model. This iterative process ensures that the model remains adaptive and effective in capturing evolving linguistic expressions within the Tamil language.

Application in Diverse Domains

The final step involves applying the SVM-RF model in practical domains such as customer feedback analysis and social media monitoring. The adaptability of the model to both written and spoken language makes it suitable for various applications where understanding sentiments and emotions is crucial. The model's outputs contribute to a deeper understanding of the emotional landscape within the Tamil-speaking community.

REFERENCES

[1] David Collins, Alan Deck, Myra McCrickard "Computer Aided Instruction: A Study Of Student Evaluations And Academic Performance", Journal of College Teaching & Learning – November 2008 , Volume 5, Number 11

[2] Bellomo, T. (2009, April)." Morphological analysis and vocabulary development: Critical criteria." Reading Matrix, 9(1),44-55: http://www.readingmatrix.com/articles/bellomo/article.pdf

[3] Joakim Nivre, "Dependency Grammar and Dependency Parsing"

[4] http://stp.lingfil.uu.se/~nivre/docs/05133.pdf

[5] Dhanalakshmi V., Padmavathy P., Anand Kumar M., Soman K.P., Rajendran S., "Chunker for Tamil," artcom, pp.436-438, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009, IEEE Press,doi: 10.1109/ARTCom.2009.191

[6] Dhanalakshmi V., Anand Kumar M., Rekha R.U., Arun Kumar C., Soman K.P., Rajendran S.,"Morphological Analyzer for Agglutinative Languages Using Machine Learning Approaches," artcom, pp.433-435, 2009 International Conference on Advances in Recent Technologies in Communication and Computing, 2009,IEEE Press,doi: 10.1109/ARTCom.2009.184

[7] Dhanalakshmi V, Anandkumar M, Vijaya M.S, Loganathan R, Soman K.P, Rajendran S, "Tamil Part-of-Speech tagger based on SVMTool", In Proceedings of the COLIPS International Conference on natural language processing(IALP), Chiang Mai, Thailand. 2008.

[8] Jes´us Gim´enez and Llu´ıs M`arquez.(2004) "SVMTool: A general pos tagger generator based on support vector machines". In Proceedings of the 4th LREC Conference, 2004.

[9] Lafferty J, McCallum A, Pereira F. 2001. "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". Proceedings of ICML: 282-289.

[10] Anand Kumar M, Dhanalakshmi V, Soman K.P and Rajendran S. "A Sequence Labeling Approach to Morphological Analyzer for Tamil Language", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 06, 2010, 2201-2208. [11] Anand Kumar M, Dhanalakshmi V, Rekha R U, Soman K.P and Rajendran S. Article: "A Novel Data Driven Algorithm for Tamil Morphological Generator", International Journal of Computer Applications 6(12):52–56, September 2010.

TAMIL LANGUAGE TECHNOLOGIES FOR EDUCATION AND E-GOVERNANCE

IoT Integrated Dairy Farmers Support Digital Solutions through Tamil Language Interfaced "Milk Productivity Improvement Technology Platform (Milk Pit)"

Dr. Palanisamy Selvaraj, Dr. A. Kavitha, Dr. S. Pravin Kumar and Dr. Vijay Jeyakumar

ABSTRACT

Many technology solutions are deployed in leading nations, while it is not so in India. Adoption of them is difficult as they will be in English and our Farmers cannot read and understand English based processes. Rural Communities of Tamil Nadu needs advanced technology in understandable language. Our approach to Tamil Nadu is developing and deploying Dairy Farmers Support Digital Solutions through Tamil Language Integrated "Milk Productivity Improvement Technology Platform". IoT Technology is used here to connect with dairy farmers who produce milk and sell it to the State's Milk Producers Co-operative Federation. Tamil Language interfaced digital solutions are the only options. Capture and monitoring the data of milk constituents of each cow, by veterinary medical experts who specializes in Dairy Animal Production Medicine, helped in early detection of milk yield reductions and helped to prevent losses. Each farmer will get immediate alerts in Tamil language, so as to help them improve production practices & health care.

Dr. Palanisamy Selvaraj
Lead PI, ICAR-NASF Artificial Intelligence & IoT SmartVet Ecosystem Project,
Dept. of Veterinary Clinical Medicine, Madras Veterinary College, Chennai and
Professor and Head, Veterinary University Peripheral Hospital,
Tamil Nadu Veterinary and Animal Sciences University
Madhavaram, Chennai.
E-mail: drdvmselvaraj@gmail.com
&
Dr. A. Kavitha, Dr. S. Pravin Kumar and Dr. Vijay Jeyakumar
Department of Biomedical Engineering,
Sri Sivasubramaniya Nadar College of Engineering,
Kalavakkam.

INTRODUCTION

Tamil Nadu is one of the leading milk producer state in India. To reach the top positions, some more interventions are needed both at grass root levels and throughout the dairy value chain. Each farmer needs to improve their cow's / buffalo's milk productivity and this requires inputs of modern knowledge, resources and practices. Many technology solutions are deployed in leading nations, while it is not so in India. Adoption of these technologies at our villages is difficult as they will be in English and majority of our farmers cannot read and understand English based processes. Rural Communities of Tamil Nadu needs advanced technology in understandable language and in their mother tongue – The Tamil. This paper presents the features of the Tamil Language Interfaced "Milk Productivity Improvement Technology Platform" – christened as "MILK PIT".

CURRENT STATUS AND LITERATURE

With India being the World's No.1 Milk Producer, our milk productivity per cow is very low. To achieve higher milk productivity per cow, our farmers need many advanced technologies. Besides technology development, their translation to rural applications is the mainstay to achieve a success. All the current technologies are existing only in English based interface. Hence it requires the services of an English literate person. The IoT Technologies developed and deployed in livestock sector so far is not exception and all of them were English based. AI and IoT based technologies are increasing used in milk production monitoring (Vishniakou and Zhifeng, 2022)

In some of the software interfaces developed for such farmer centric technologies does not have Tamil as an Option and our farmers are not able to understand any information from these machines. No literature exist in the public domain about the use of Tamil as interface in technology enabled service delivery to Livestock Farmers.

Hindi, Telugu, Kannada were tried by some commercial entities in the Milk Collection Software with varied success. Still no efforts were there on to use Tamil language 9nterface in Milk quality analyzers. For

the first time, we are developing Tamil language based user interface in the IoT enabled Milk Testers.

METHODOLOGY

Our approach to Tamil Nadu is developing and deploying Dairy Farmers Support Digital Solutions through Tamil Language Integrated “Milk Productivity Improvement Technology Platform”. IoT Technology is used here to connect with dairy farmers who produce milk and sell it to the State’s Milk Producers Co-operative Federation. As many of them were not educated nor English speakers, Tamil Language interfaced digital solutions are the only options.

The Milk is analyzed by Milk Tester Machines available at Village Milk Societies and price of milk is fixed based on its fat content. During this analytical process, many milk constituents are measured like fat, solids not fat, protein, etc. and all these data are simply discarded once the price is fixed. These data are very vital and essential for improving milk productivity of each cow.

Unfortunately currently these data are not at all utilized and hence no attempts are being made to improve milk productivity. Monitoring the milk constituents of each cow, by veterinary medical experts who specializes in Dairy Animal Production Medicine, early detection of milk yield reductions can be identified and losses can be prevented. Each farmer will be sent immediate alerts in Tamil language, so as to help them improve production practices & health care. Current technology prototype was successful and was validated under field conditions.

Tamil Language Interface Development

In our villages farmers milk the cattle and sell majority of their milk to the village milk co-operative society. In the Milk Society, then milk is tested using an analyser and it gives out a print out or display of some parameters. These are the levels of some constituents present in the milk and using some of them, the milk price is fixed. The remaining data output is just left out or trashed. If an IoT module access these discarded data and with the help of Large Language Models (LLMs), we are able to develop farmer centric information empowerment to increase the milk productivity.

LLMs are used as a method of answering any of the farmer's questions. Firstly, the correct LLM is evolved and deployed for this purpose. This LLM shall be able to handle the native language of our farmers - Tamil. In its next stage, the LLMs are given data about the intricacies of dairy animal farming so that all of the farmer’s queries can be answered. This information must be taken into account productivity and how friendly it is to the environment and give information accordingly so the farmer can choose what is best for them.

There have been a few models that aimed to use different vernacular languages like Tamil for their training dataset. Initially, the integration of Automatic Speech Recognition (ASR) with ConvLSTM Networks represented a groundbreaking leap in local language detection. The speech was recognized using the ASR and this served as the database for the model. The model employed ConvLSTM which combined the capabilities of CNN and a Long Short Term Memory model. However, constructing a Language Model (LLM) for Tamil, especially for farmers and rural communities, presents challenges. For other languages too, it presents challenges due to its diverse local and geographical dialects. As a result, leveraging a semi-supervised speech corpus significantly improved outcomes for the complexity inherent in the language's variations.

Recently, the llama2 model was used to address the limitation of underrepresentation of Tamil in language models like ChatGPT by adding 16,000 Tamil tokens to enhance text generation and comprehension. It used LoRA(Long Range) methodology and tailored datasets, to achieve performance improvements in Tamil language tasks. These research aimed to encourage further advancements and innovations in language modeling by making the models, datasets and code openly available. Thus, with further improvements in optimizing LLMs for different languages such as Tamil, they can be incorporated into various use cases. The datasets can be improved and thus used for fine tuning the parameters. Farmers, especially, would benefit greatly from the implementation of LLMs to improve their productivity and performance.

Mobile Phone based Tamil Interface for IoT Milk Tester:

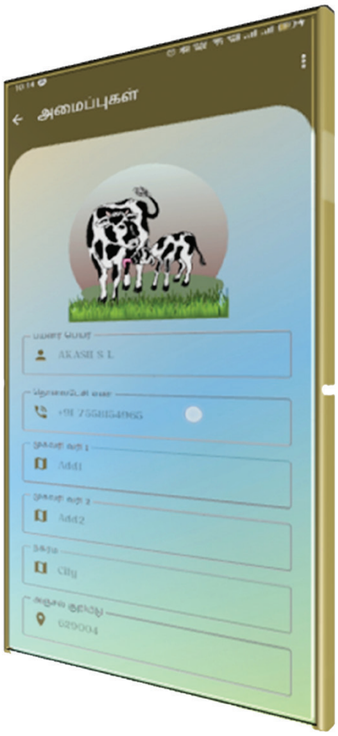
With emerging knowledge, it is possible to develop an app that can show nothing but Tamil. The disadvantage here in we can't code fully in Tamil yet. Rather, we don't code in English or Hindi or Spanish, etc. Only Java, Python, C+ etc. are the platforms used to develop coding and these languages are currently in English script only. It will be a long time before they come up in any other script.

Tamil Farers Friendly User Interface:

To facilitate every family member of our rural families and farmers understand the process and enable its operation by ever one. They will get to know in Tamil language, the results and the advices to rectify

mistakes. This enables the farmers themselves to undertake earliest intervention and there by restore production. Most importantly these Tamil language based guidance's for early farm gate early interventions helps to prevent escalation of further losses and severity of ill health and related issues, all of which reduces the associated expenditure costs and saves the farmers the money and ensures continuity of production. It has bilingual modality with English, so as to trouble shoot the issues.

Cloud & Fog Computing based Support using Tamil Interface for Livestock Farmers

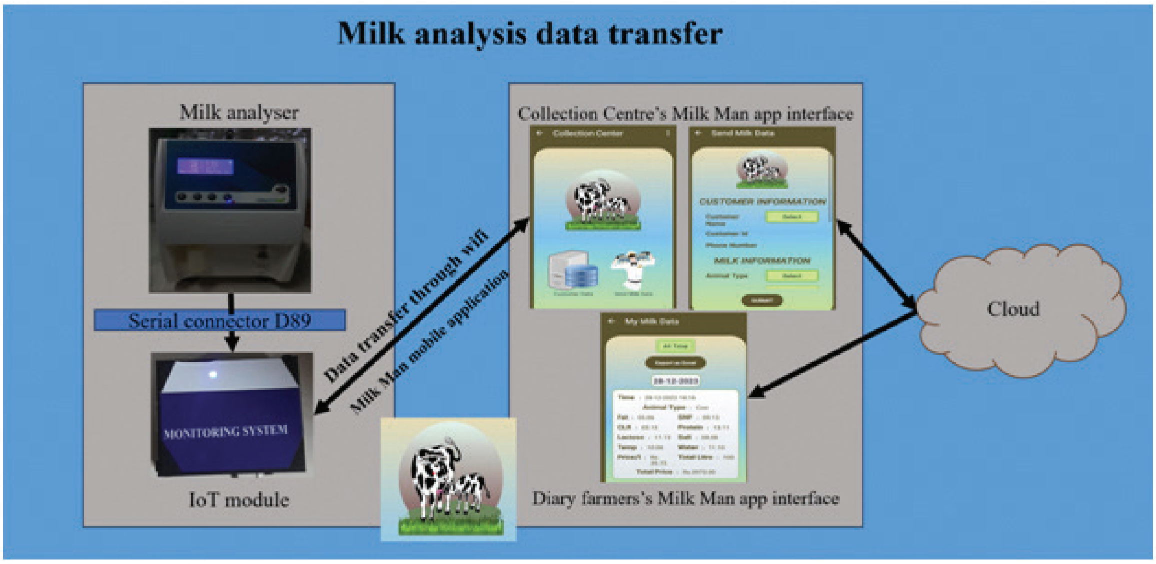




Technology behind this Farmer Intervention Centric Milk Tester

Advanced Technologies are mainstay in manufacturing and related sectors. Can any of these technologies can be directly be deployed for usage by illiterate farmers or rural families who does not know English? It is not possible as of now. When we can develop Language Technologies and Language Interface Technologies for emerging modern tools, then technologies like Internet of Things (IoT), fog computing, cloud computing, and data-driven techniques can be fine tweaked and deployed and all together offer a great opportunity for not only verticals such as dairy industry, but also our rural milk producers with just one or two dairy cows per rural household to increase productivity by getting actionable insights in their language – the Tamil, to improve dairy farming practices and milch animal health care, thereby increasing the efficiency and yield.

Our approach here is, developing a cloud and fog computing–assisted end-to-end IoT platform for usage at village milk cooperative societies and milk collection points for milk analysis and there by using the milk constituent data for health monitoring in our predominantly rural and extremely small scale milk producers as well as for usage in rest of the dairy farming scenario in the state.



RESULTS AND DISCUSSION

Technology is essential for development of a society and a state. Without people understanding the technology, there is no scope for development. Unless it is transferred in majority of the people's preferred language, such technology transfers cannot be achieved. Societal transformation needs language centric technology approaches.

With Tamil Nadu leading in many domains, the emerging advanced technologies also needs to be integrated with language interfaces. Livestock Farming of Tami Nadu is at the cross roads of transformation. It plays a major role in the development of rural communities of Tamil Nadu. Tamil Language based technology interventions are our goals and presenting here with one such technology here.

Our approach is developing low cost module so as to make it field adaptable with high success rates. Commercial entities are selling IoT Milk Analyzers at higher costs, which may not be practicable for the resource constrained village milk co-operative societies. Beside, these societies have already older versions of machines and may not be interested to get new ones or have resources to invest in new gadgets. Hence converting existing old/conventional milk testers would be an easy option and hence development of an IoT Connector Module becomes practicable approach here. It also resulted in cost cutting and easy translation for field applications.

Sensors that measure the milk components and these are combined to create compact and versatile system that characterizes the quality of milk into data and these are finally showed on alphanumeric showcase screen. Conjointly with the assistance of IoT tools, the data on milk can be sent to the producer and end users and the intermediaries in between. It helps the stake holders, policy makers and the governments. With the Tamil Language interface all becomes much easier for every one the dairy value chain can easily understand and it becomes an actionable data.



Advantage of the IoT technology is that it allows the devices and objects to be sensed or controlled remotely across existing network infrastructure, creating opportunities for more direct integration of the physical world into computer-based systems, and resulting in improved efficiency, accuracy and economic benefit in addition to reduced human interventions. When IoT is augmented with sensors and actuators, it becomes an instance of a general class of cyber-physical systems.

Milk testing instruments are made to help accurately analyze the properties or characteristics of raw cow and buffalo milk such as fat, SNF, added water, density, protein, lactose. Unfortunately all of these data are not utilized in every day farming practice, while phenomenal scope exists to improve farming practices and farmers' incomes. IoT integration with such testers provides not only the measurement of mixed milk components, of both cows and buffaloes, but the effective analysis of the data and deriving very actionable insights from those data.

In fact in many of the private dairy organizations, smart data processing based milk collection and analysis systems were already being deployed while it is not so in the government run cooperative milk sector units. Such smart solutions helps efficiently manage the entire dairy supply chain, making the process faster, more accurate and transparent. While it not only receives, processes and transmits data, but it also supports to improve the conventional village-level milk collection systems which is used by milk producers and milk contractors.

For the dairy value chain, edge computing and edge devices are going to be the game changes. We already use some devices that do edge computing every day—like smart speakers, watches and phones – devices which are locally collecting and processing data while touching the physical world. Internet of Things (IoT) devices, point of sales (POS) systems, robots, vehicles and sensors can all be edge devices—if they compute locally and talk to the cloud. When these are deployed innovatively, all of it will help to improve farmer's income, especially the farmers are empowered with data driven actionable information in their mother tongue – Tamil.

Total Milk Information in Tamil for Farmers to Know and Act





CONCLUSION:

Emerging Digital Technologies like Artificial Intelligence holds great economic, social, medical, security and environmental promises. Unfortunately they are not tapped for the growth of livestock sector in our state and that exactly is what the Tamil Nadu state government agencies need to explore and exploit. Our State's Livestock sector shall aim to harness the power of AI for the public good while making it ethically compatible with human values.

Our efforts and approach through developing and deploying Dairy Farmers Support Digital Solutions through Tamil Language Integrated "Milk Productivity Improvement Technology Platform" is fruitful. The developed prototype using IoT Technology is found to be successful and is being used to connect with dairy farmers. Tamil Language interfaced digital solutions capturing and monitoring the data of milk constituents of each cow, helped farmers. More importantly it helped veterinary medical experts who specializes in Dairy Animal Production Medicine, towards developing strategies for early detection of milk yield reductions and helped to prevent losses. Farmers were able to get immediate alerts in Tamil language, so as to help them improve production practices & health care.

FUTURE POTENTIALS:

Government of Tamil Nadu had many initiatives towards bringing in Tamil Interface with all service delivery platforms. The TNeGA has already developed a Tamil chatbot, named "Anil" , using Natural Language Processing and Artificial Intelligence Technologies. This Tamil chatbot is helping to guide and advise people by answering their queries and guide them to obtain government services like obtaining nativity certificates, income certificates, government certificates etc. When such Tamil, based chatbots are integrated in to Farmers Technology Platforms like this "Milk PIT", it will further help improve farmer's day to day abilities in solving many of their problems and increase their revenue potentials, and overall prosperity

REFERENCES:

Vishniakou U.A., Zhifeng H. Development and Optimization of the Internet of Things Network for Product Quality Monitoring. Doklady BGUIR. 2022;20 (4):80-87. (In Russ.) https://doi.org/10.35596/1729-7648-2022-20-4-80-7

Development of a Tamil Handwriting App that Offers a Guided Approach for Children to Learn, Practice and Enjoy Tamil Handwriting

Khasturi Ramalingam, Muthu Nedumaran

ABSTRACT

Handwriting is and essential skill, especially among children. Writing by hand requires greater cognitive involvement than typing. Stroking each letter engages various neural pathways in the brain. These trigger regions linked to language, memory, and motor abilities. Proficient handwriting helps with effective communication and clear expression of ideas which in turn provides a foundation for lifelong learning and academic success. In Malaysia, children learn Tamil alongside two other languages: Malay and English which are written in Latin. The complex forms of Tamil letters, relative to Latin, and the lack of resources to practice is becoming a demotivating factor for children to learn and enjoy Tamil handwriting. This research aims to help fill that gap through the creation of a tablet-based application, with seven key features, that will guide children using simple data models. The application will show animated stroke movements of each letter, let children follow along the movements, and assess the accuracy of the final letter forms written by the children. To study the effectiveness of the approach, 20 children aged 7 were selected as sample from 4 different schools. Qualitative and quantitative methods were used to collect data and analyse the outcome. In conclusion, this research presents a pioneering method that combines machine learning techniques and motivational strategies to assist children in honing their Tamil handwriting skills. The integration of technology not only enhances learning outcomes but also instils a sense of enthusiasm, making the process of learning Tamil handwriting enjoyable and rewarding for young learners.

Khasturi Ramalingam, Muthu Nedumaran
SJK(T) Masai, Johor, Malaysia
Murasu Systems Sdn Bhd, Selangor, Malaysia

1. INTRODUCTION

Research by others have shown that Tamil students in Malaysia encounter hurdles in mastering Tamil handwriting due to inadequate resources, limited guidance, and inconsistent practice (Rudrapathy & Rudrapathy, 2022). The difficulties they face include the complexity in Tamil letter forms, inability to write fluently and smoothly as they do when they write English or Malay with Latin alphabets. These difficulties make Tamil handwriting practice uninteresting (Winskel, 2020). It takes less strokes and curves to write Latin alphabets compared to Tamil. However, the many curvatures in Tamil letters can add to the beauty of the forms with correct stroke movements (Nedumaran, 2018).

The aim of this research is to develop a tablet-based Tamil handwriting app, with seven key features, that offers a guided approach (THAGA) for children to learn, practice and enjoy Tamil handwriting with a digital pen.

THAGA needs to address the above concerns. At the same time, it should synchronise with the tools currently used by the children in schools, like worksheets and activity books.

The first thing we did was to discuss and arrive at a list of seven key features we want THAGA to have. Using these key features as the requirements, we evaluated nine existing handwriting apps for Tamil language that are available in Apple's App Store and Google Play for tablets. Among the nine, three were available on both platforms. While some of the apps had some of the features, none of them had all the seven features we enlisted. Since none of the apps met all features in our list, we went ahead and built a prototype to implement our features with a minimal set of data for selected letters. The purpose of this prototype was to study the effectiveness of the guided approach with six of the seven features, before developing the full-fledged app.

With this prototype, we went ahead to conduct our second phase of the study: to evaluate the effectiveness of the guided approach used in THAGA to help improve the ability of children to write the selected letters in the correct form. The study was also expanded to ascertain if the children learnt, practiced, and enjoyed writing the

Tamil letters with THAGA. We also wanted to ensure 21st century values such as self-directed learning, ICT literacy and collaborative learning can be applied to the child or a group of children indirectly.

2. SEVEN KEY FEATURES OF THAGA

We analysed the current tools like worksheets and activity books used in schools to teach Tamil handwriting. Since THAGA is targeted as a supplementary tool, alongside current tools, the feature set must complement the current tools and not isolate itself with a completely different approach to teaching Tamil handwriting. It can, however, improve the current approach with the use of technology and machine learning capabilities. Based on this principle, we arrived at these seven key features:

i. The letter forms used as templates in the tool should closely match a person's handwriting as opposed to a letter from a typeface designed by a type-designer.

ii. The tool should take advantage of writing instruments like digital pens or pencils to closely match real world handwriting experience with a regular pen or pencil.

iii. The writing area should provide similar metrics lines as provided in paper worksheets and activity books.

iv. Include animated movements of the correct stroke directions, instead of arrows along the lines which sometimes can appear congested and cause confusion.

v. When assessing the letter drawn by the child, compare the directions of all strokes against the template form, instead of only the overall shape of the final form. This will ensure that the letter is written in the correct direction as shown in the animation in (iv).

vi. As the child practices drawing the letters, learn the correct forms drawn by the child and move the template to those learnt forms instead of sticking with the original template. This will indirectly encourage the child to develop a unique writing style and score higher points when new drawings of the letters are assessed.

vii. When the practice has covered all the required letters in Tamil, provide the ability to export the final forms of the drawn letters into a font that can be used in all applications.

3. LIMITATIONS IN CURRENTLY AVAILABLE APPS

None of the apps we evaluated included all the seven key features. The most notable absence that we saw as critical limitations were features (i) and (v). All apps used an embedded Tamil font or the resident Tamil font in the system, which were designed by type designers. They did not appear as a naturally handwritten letter. The resident fonts were not designed for this intent.

Also, when tracing the letters, the user was forced to stay within the stroke boundaries, even if the strokes were moving in the right direction. In other words, the assessment was done with strict compliance to the exact strokes of the template letter and not the overall stroke directions and form of the letter.

4 DEVELOPMENT & IMPLEMENTATION OF THAGA PROTOTYPE

We wanted to study the effectiveness of the first six of the seven features before beginning full-fledged development of THAGA. For this purpose, we developed a prototype that allowed children, who will be the actual users of the app, to learn and practice a few letters. We interviewed some teachers who teach children Tamil handwriting to get their perception on which letters needed the most attention. They explained that children generally learn simple letters like ட, ப and ம quite easily. However, they have difficulties writing other complex letters. Based on this feedback, we chose four letters where most children had difficulty writing: இ, த, ழ, and ஐ.

The seventh feature in our list required all letter forms to be available in order to create a font. As such, we dropped this in the prototype and kept it for the full-fledged app.

The prototype app was built for iPads with Apple Pencil, running iPadOS 15.0 and later. We chose this platform as it had readily available frameworks like SwiftUI and PencilKit with which we could quickly build the user interface and add stroking features to draw letters. With this setup, we started adding the first six of the seven key features.

4.1 Template letters

The template letters were drawn in the prototype app itself. The strokes, which included the start point, end point and direction information, were captured as drawings. These drawings were saved into binary files and used as the data model to access the letters drawn by the children. We used the available functions in iPadOS that allowed developers to capture the drawing data. These functions are published in the PencilKit documentation (Apple Developer, 2024).

4.2 Apple pencil

Apple's iPadOS includes excellent integration with the ApplePencil through the PencilKit framework. The latest version of this framework includes a monoline inking tool that draws strokes with uniform thickness irrespective of the speed and direction of the strokes. This closely emulates a regular pencil or pen. In earlier frameworks, this inking tool was absent. However, we could adjust the thickness to make it uniform after the stroke is drawn. In either case, the second key feature can be met with this framework.

4.3 Metrics lines

We drew 4 horizontal lines in the app's writing area: baseline, letter height, ascender, and descender. This follows the lines drawn in paper worksheets and activity books for Tamil handwriting used in Malaysia.

4.4 Animating strokes

Since the prototypes are stored as PencilKit drawing data, which in turn contain vector data, animating them was made possible with SwiftUI's animation frameworks (Apple Inc, 2023). The speed at which the strokes are drawn can also be adjusted. The user can either choose to keep the template visible while the strokes are animated or just show the drawing of the strokes alone.

4.5 Assessing user drawn strokes

When users finish writing a letter the stokes of the written letter is compared against the stokes for that letter in the data model. Comparison is done using Fréchet distance. Fréchet distance measures the similarity between curves that takes into account the location and ordering of the points along curves (Figueira, 2020). By adjusting the acceptable distance, we can allow flexibility in scoring. For the prototype, we set an arbitrary minimum distance and adjusted it based on a few inputs from children. This is to ensure that the assessment is neither too restrictive nor too loose. The main aim is to get children to write the letter in the correct order and shape. It need not perfectly match the template form in the model. Also, ability to score easily will help motivate the children to keep writing and this in turn will give us more strokes for the learning as explained in 4.6.

4.6 Learning new strokes

When the children complete writing a letter and if the assessment of the strokes in the letter based on Fréchet distance hit a score higher than 90, the drawing data of that newly written letter is added to the data model. This will serve as additional data when evaluating the score of future strokes for the same letter from the same child. As more and more such data are added to the model, the tool will evaluate future writings to match closer to the child's writing instead of the template. This will let children keep hitting higher scores in the app, again serving as a motivation to make them keep writing.

5. EVALUATION

Before embarking on the research process to evaluate the features of THAGA with sampled users, the prototype was shown to three experts from three different universities in Malaysia: UTM, UPM and UM, to validate the features for use in schools. The three experts reviewed and asserted that the features were in line with the aim of THAGA, which is to provide a guided approach for children to learn, practice and enjoy Tamil handwriting with a digital pen.

With the experts' clearance and the six features coded in, the prototype app was used to evaluate the following three areas: 1) effectiveness of the guided approach to teach the letter forms, 2) the perception from the teachers about the features of this app, and 3) if the students showed enthusiasm in learning and enjoyed writing the selected letters.

This research process consisted of the following four steps 1) plan, 2) implement, 3) observe, and 4) reflect. These four steps were adopted based on the model developed by Kemmis and McTaggart (Kemmis et al., 2014).

The research was guided by four teachers, one each from four different schools. The teachers selected five children who were students, aged seven, from each school with written consent from the school and parents. These students were selected based on their low achievement in writing. We also prepared three iPads with THAGA prototype pre-installed for the children. Classes were planned for five sessions in the month of November 2023. Each session lasted for a period of 30 minutes. A pre-test was given by the teachers before the first session started. In the pre-test, the students were asked to write these four letters on a lined paper with a pencil.

In the first four sessions the students wrote on the iPad running the THAGA prototype. In the final session, post-tests were conducted to measure progress in handwriting skills for the same four letters.

6. OUTCOME OF EVALUATION

The outcome of the evaluation was positive for all the three areas. The guided approach was effective in improving the handwriting among the students, the perception from the teachers about the features was positive and the students showed enthusiasm in learning to write the Tamil letter forms.

6.1 Improvements in handwriting

Figure 1 shows the percentage comparison analysis between pre-test and post-test for handwriting based on adapted rubric scores from (Ganesan & Abu Bakar, 2023; John & Renumol, 2022). The scores were collected from the pre- and post-tests. The graph clearly shows that all the 20 study participants increased their scores in the post-test, which was conducted after using THAGA to learn and practice the four chosen letters: இ, த, ழ, and ஐ. The improvements in scores were more than 50% across all the letters. Writing of letter இ improved by 62%, த improved by 50.8%, ழ by 57.2%, ஐ by 52%.

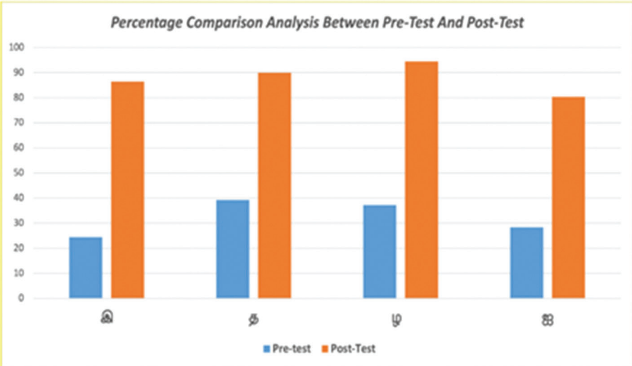


Figure 1: Analysis between pre-test and post-test .

Overall, the results of the analysis showed that the use of the THAGA helped improve the mastery of writing the selected letters, which were relatively difficult, among the 20 study participants.

6.2 Teachers' perception of THAGA's features

Teachers' perceptions of THAGA's features were assessed using a series of interview questions. Overall, teachers' feedback indicated an increased interest in writing among students after using the THAGA prototype. Additionally, the teachers affirmed that the features are suitable and user-friendly, specifically for children. They liked the way the app presented and animated the template letters which clearly showed the start point, the stroke direction and the end point. This was seen suitable for beginners to write the letters correctly, which is a problem they have been facing. Children write the letters in the correct form but often in the wrong direction. Teachers also mentioned that ApplePencil on the iPad gave the children the feeling of using a regular pencil. They became accustomed with the tool very quickly and started using it immediately. The teachers were also happy with the 4 lines feature as it followed the lines in worksheets and activity books for Tamil handwriting used in the classroom. Finally, they liked the way THAGA evaluated the letter forms written by the children. Although it was internally using, Fréchet distance, a method that they could not understand, they found the result to be extremely interesting. Especially when it took into consideration the correct stroke paths of the letters. When the children wrote the letters from another direction, the assessment clearly showed that it was wrong.

6.3 Students' enthusiasm

Teachers reported that students who participated in the study thoroughly enjoyed writing the letters இ, த, ழ, and ஐ using THAGA. Additionally, the teachers noted that the students found it easy and straightforward to write these letters after practicing with the prototype app. Furthermore, the teachers observed that the students were largely self-directed and engaged in collaborative efforts with their peers in the classroom.

7. CONSIDERATIONS FOR THE FULL APP

The evaluation exercise conducted with the prototype app provided us with the confidence that the seven key features we enlisted for THAGA will effectively fulfil our goal of creating a tablet-based Tamil handwriting app that offers a guided approach for children to learn, practice, and enjoy Tamil handwriting using a digital pen.

The prototype app only focused on handwriting practice on இ, த, ழ, and ஐ. The full app will need a data model that contains template strokes for all the forms required to write Tamil. These forms will include vowels, consonants, vowel-signs and the aytham. For beginners, these forms can be introduced in groups that are organised by simplicity and shape of the forms.

Besides letters, THAGA can also include exercises for children to write words and sentences. The application can include a large wordlist with the ability to extract words that contain only selected letters through a regular expression search. Such a list and the search expressions exist in the open sourced Hibizcus project by Muthu Nedumaran (Nedumaran, 2023).

7.1 Font creation

The seventh key feature we enlisted for THAGA was the ability to export the letter forms written by the user as a complete working Unicode font, with all OpenType shaping logic included in it. This will serve as a great motivation for the children to use their own font when they learn touch typing in Tamil in the future.

7.2 Other platforms

The prototype depended heavily on frameworks available in Apple's operating systems, especially iOS and iPadOS. Frameworks like PencilKit greatly simplified stylus and finger-based drawing on Apple's devices. To create a similar user experience on

other platforms like Android and Windows, similar frameworks need to be developed.

8. CONCLUSION

In the prototype of THAGA, only 4 letters were chosen to evaluate the effectiveness of the approach with six of the seven key features. As described in section 6, the children and the teachers found the approach to be effective. Using a guided approach with templates that match strokes, curves and their direction proved to be very effective for children to learn Tamil handwriting. The results of the evaluation further strengthen the key features that we had identified for THAGA.

The final app need not be confined to seven-year-old children, whom we chose as the sample to do the evaluation. It can be used to as a tool for any children in any age group who want to learn or improve their Tamil handwriting which in turn brings in the benefits of writing by hand. The app can also be offered to non-Tamil speaking children or adults across the world who are interested in learning Tamil handwriting.

REFERENCES

- Developer, A. (2024). PencilKit: Capture Touch and Apple Pencil Input as a Drawing, and Display That Content From Your App. Apple Inc.
- Figueira, J. P. (2020). Culculating the Discrete Frechet Distance Between Curves. Medium.Com, 1(1), 1–10.
- Ganesan, D., & Abu Bakar, K. (2023). Penggunaan Kaedah Permainan 'Maze' Bagi Mengatasi Masalah Menulis Huruf Besar Berbentuk Lengkok C, O, S dan U. Malaysian Journal of Social Sciences and Humanities (MJSSH), 8(10), 1–12. https://doi.org/10.47405/mjssh.v8i10.2441
- John, S., & Renumol, V. G. (2022). Design and Development of an Android App (HanDex) to Enhance Hand Dexterity in Children with Poor Handwriting. IEEE Access, 10, 48973–48993. https://doi.org/10.1109/ACCESS.2022.3172330
- Kemmis, S., McTaggart, R., & Nixon, R. (2014). The Action Research Planner. The Action Research Planner. https://doi.org/10.1007/978-981-560-67-2
- Nedumaran, M. (2018). Exploring A Cursive Design for a Mixed Script Typeface. Beauty, Form and Function in Typography, 1–11.
- Nedumaran, M. (2023). https://github.com/murasu/Hibizcus. https:// hibizcus.com. https://github.com/murasu/Hibizcus
- Rudrapathy, C. S., & Rudrapathy, T. (2022). Reading and writing skills performance level of students learning tamil as a second language in national primary school. Muallim Journal of Social Science and Humanities, 6(1), 46–54. https://doi.org/https://doi.org/10.33306/mjssh/176
- Winskel, H. (2020). Learning to Read in Multilingual Malaysia: A Focus on Bahasa Melayu, Tamil and Chinese. GEMA Online Journal of Language Studies, 20(1), 1–15. https://doi.org/10.17576/gema-2020-2001-01

டிஜிட்டல் உலகில் தமிழ் வாசிப்பு: நேற்று, இன்று, நாளை

பா. ராகவன்

ஆய்வுச்சுருக்கம்

பதினாறாம் நூற்றாண்டில் வெளியான முதல் தமிழ் நூலிலிருந்து இன்று வந்திருக்கும் கிண்டில் செல்பேசி செயலிகள் வரை தமிழ் நூல்களும் வாசகர்களும் எப்படி வளர்ந்திருக்கிறார்கள் என்பதை இக்கட்டுரை ஆராய்கிறது.

அடுத்து வந்துகொண்டிருக்கும் ஈனும் செயற்கை நுண்ணறிவு காலத்தில் எழுத்தாளர் என்பவர் யார்? அவரது பணி என்னவாக இருக்கும்? இனி வரும் காலங்களில் கணினியும் செயற்கை நுண்ணறிவும் இல்லாமல் தமிழ் எழுத்தும் வாசிப்பும் இயங்க முடியாது என்கிற சூழ்நிலையில் தமிழ் வாசிப்பின் அடுத்தக் கட்டம் என்னவாக இருக்கப் போகிறது என்றும் இக்கட்டுரை ஆராய்கிறது.

பா. ராகவன்
ஆசிரியர், மெட்ராஸ் பேப்பர்

வாசிப்பு: வந்த வழியும் வளரும் வழியும்

முப்பத்தைந்து ஆண்டுகளாக எழுதிக்கொண்டிருக்கிறேன், எழுத்துத் துறையில் மட்டும்தான் இயங்கிக்கொண்டிருக்கிறேன். இந்த முப்பத்தைந்து ஆண்டுகளில் இந்தத் துறை சார்ந்து பேசப்பட்ட கருத்துகளுள் முதன்மையானது, 'வாசிப்புப் பழக்கம் குறைந்துவிட்டது' என்பதுதான். பதிப்புத் துறையிலும் சரி, பத்திரிகைத் துறையிலும் சரி. யாரைக் கேட்டாலும் தயங்காமல் இதனைச் சொல்வார்கள். இரண்டாயிரத்துப் பத்தொன்பதாம் ஆண்டுக்கு முன்புவரை இக்கூற்றை உறுதி செய்வதற்குப் புள்ளிவிவரங்களைத் தேடி எடுக்க வேண்டும். கோவிட் பெருந்தொற்றுக்குப் பிறகு நிறுத்தப்பட்ட பத்திரிகைகள், முடங்கிப்போன பதிப்பகங்களைச் சுட்டிக்காட்டினாலே போதும் என்ற நிலை உண்டாகியிருக்கிறது.

நல்லது. இது ஒரு சிக்கல். ஆனால் தீர்க்க முடியாத சிக்கல் அல்ல. இதன் தீர்வு, இந்தச் சிக்கலை நாம் எப்படிப் புரிந்துகொள்கிறோம் என்பதில் இருக்கிறது. பொதுவாக, தமிழ்ச் சமூகம் வாசிப்பிலிருந்து விலகிச் செல்கிறது என்று குற்றம் சாட்டுவதை விடுத்து, நாம் எப்படி இருந்தோம், எங்கிருந்து எங்கே நகர்ந்து வந்திருக்கிறோம் என்று சுய மதிப்பீடு செய்து பார்ப்பது இவ்விஷயத்தில் சரியான பலனைத் தரும்.

கிபி 1492ஆம் ஆண்டுதான் அமெரிக்கா என்ற நாடே கண்டுபிடிக்கப்படுகிறது. ஆனால் கிபி 1554 பிப்ரவரி 11ஆம் தேதி போர்ச்சுகலின் தலைநகரான லிஸ்பனில் முதல் தமிழ்ப் புத்தகம் அச்சாகிவிட்டது. இன்றைக்கு ஆங்கில லிபியில் தமிழை எழுதுவதை தங்கிலீஷ் என்கிறோம் அல்லவா? அன்றைக்குத் தமிழை லத்தீன் மொழியில் எழுதிப் பதிப்பித்தார்கள். 'தமிழ் மொழியிலும் போர்த்துகீசியத்திலும் அமைந்த திருமறைச் சிற்றேடு' [1] என்பது அந்நூலின் பெயர்.

கவனிக்க வேண்டிய இடம் இதுதான். பதினாறாம் நூற்றாண்டில் இந்த முதல் தமிழ்ப் புத்தகம் அச்சான சமயத்தில் இங்கே தமிழ்நாட்டில் செப்புப் பட்டயங்களில் அரசச் செய்திகள் வெளியாகிக்கொண்டிருந்தன. கல்வெட்டில் எழுதும் வழக்கமும் இருந்தது. அச்சு நூல்கள் வெளியாகத் தொடங்கியதால் நமது பாரம்பரியமான செப்புப் பட்டயங்களில் எழுதுவதும் கல்வெட்டில் எழுதுவதும் அருகி, இல்லாமல் போய்விட்டன என்று என்றாவது வருத்தப்பட்டிருக்கிறோமா?

இதுவேதான் இன்றைய 'வாசிப்பு அருகிவிட்டது' என்கிற வாதத்துக்கும் நாம் முன்வைக்கக்கூடிய எளிய பதில். வாசிப்பு குறையவில்லை. ஆனால் வேறு தடங்களில் விலகி முன்னேறத் தொடங்கியிருக்கிறது. தமிழ் வாசிப்பு என்பது டிஜிட்டல்மயமாகியிருக்கிறது. இணையதளங்கள், வலைப்பதிவுகள், யாஹூ, கூகுள் குழுமங்கள், ஆர்குட், பஸ் என்று தொடங்கி இன்று ட்விட்டர், ஃபேஸ்புக், இன்ஸ்டாக்ராம், த்ரெட் என்று விரிந்திருக்கிறது. மின் நூல்கள், ஒலி நூல்கள், ஒலி–ஒளி நூல்கள் ஏராளமாக வரத் தொடங்கியிருக்கின்றன. கால மாற்றத்தைத் தவிர்க்க முடியாதது போலவே, அந்தந்தக் காலகட்டத்தின் வளர்ச்சிகளை உள்வாங்கிச் செழிப்பதையும் ஒரு செம்மொழி தவிர்க்க விரும்பாது.

நாம் இம்மாற்றம் நிகழ்ந்துகொண்டிருக்கும் காலத்தில் வாழ்வதால் இரு தரப்பையும் கவனிக்க முடிகிறது. உணர்ச்சிவசப்படாமல் அலசிப் பார்க்க முடிகிறது.

இம்மாற்றம் தொண்ணூறுகளின் இறுதியில் நிகழ ஆரம்பித்தது. இணையம் அப்போது ஓர் ஆடம்பரம். நிறுவனங்களில் இருக்கும். வசதி படைத்தவர்களின் வீடுகளில் மட்டும் இருக்கும். மின்னஞ்சல் அனுப்ப கம்ப்யூட்டர் சென்டருக்குச் சென்று வந்தேன் என்று சொல்வதும் ஓர் அந்தஸ்து அறிகுறியாகப் பார்க்கப்பட்ட காலம் அது. இந்த நூற்றாண்டின் தொடக்க ஆண்டுகளில் வைஃபை தொழில்நுட்பம் இந்தியாவில் பரவ ஆரம்பித்து, சாதாரண மக்களுக்கும் இணையம் சாத்தியம் என்றான பின்பு டிஜிட்டல் வாசிப்பு வேகமெடுக்கத் தொடங்கியது. ஆங்கிலம், ஜெர்மன், பிரெஞ்ச், ஸ்பானிஷ் மொழிகளை அடுத்து அதிக வலைப்பதிவுகளைக் கொண்ட மொழியாகத் தமிழ் திகழ்ந்தது. மறைந்த தமிழ் அறிஞர் அவ்வை நடராசன் 2004 ஆம் ஆண்டு ஓர் இலக்கிய மேடையிலேயே இத்தகவலைத் தெரிவித்தார். உண்மையில், தமிழில் எழுதுவோரும் படிப்போரும் கணிசமாக அதிகரிக்கத் தொடங்கியது இரண்டாயிரமாவது ஆண்டுக்குப் பிறகுதான்.

அதற்கு முன்னால் எண்ணிக்கையாகச் சொல்லப்பட்ட அதிகபட்ச சாதனை எதுவென்று ஒரு கணம் சிந்தித்துப் பாருங்கள். குமுதம் வார இதழ் ஒரு குறிப்பிட்ட காலகட்டத்தில் ஆறு லட்சம் பிரதிகள் விற்றதை மட்டுமே நினைவுகூர முடியும். நாளிதழ்களிலேயே மிக அதிக விற்பனை காணும் தினத்தந்தி தனித்தனியே பதினாறு பிராந்தியங்களில் அச்சிடப்பட்டு வெளியாகிறது. இந்நாளிதழின் அதிகபட்ச விற்பனையாகச் சுட்டிக்காட்டப்படுவது, 2015 ஆம் ஆண்டின் இரண்டாம் பகுதியில் இது எட்டிய பதினேழு லட்சம் என்கிற எண்ணிக்கை. [2] தமிழ் அறிந்த மக்களின் எண்ணிக்கையோடு இந்த எண்ணை ஒப்பிடக்கூட முடியாது. உலகில், தமிழைத் தாய்மொழியாகக் கொண்டு பேசுவோரின் எண்ணிக்கை எண்பத்தொன்பது கோடி. அதனை நினைவுகூர்ந்தால் மேற்சொன்ன எண்ணிக்கை ஒன்றுமே இல்லை என்பது விளங்கிவிடும்.

ஒரு வெகுஜன வாரப் பத்திரிகை, வெகுஜன நாளிதழின் அதிகபட்ச எண்ணிக்கையே தமிழில் இதுவாகத்தான் இருந்திருக்கிறது என்னும்போது புத்தகங்களின் விற்பனை எப்படி இருக்கும்?

ஆண்டுக்குத் தோராயமாகப் பதினைந்தாயிரம் தமிழ் நூல்கள் [4] வெளியாகின்றன. இந்திய அளவில் இந்தி மொழிக்கு அடுத்தபடியாக அதிக எண்ணிக்கையில் புத்தகங்கள் வெளிவருவது தமிழில்தான். இதர மூன்று தென்னிந்திய மொழிகளில் வெளியாகும் நூல்களின் எண்ணிக்கை இதில் பாதியளவுகூட இல்லை. ஆனால் புத்தக வாசகர்கள் என்று பார்த்தால் தமிழில் அதிகபட்சம் இரண்டு லட்சம் பேரைச் சொல்ல முடியும். இந்த எண்ணிக்கைக்கு ஆதாரப் புள்ளிவிவரமாக ஏதுமில்லை. ஆனால் பத்தாண்டுக் காலம் தமிழின் முன்னணி பதிப்பு நிறுவனம் ஒன்றின் தலைமை ஆசிரியராகப் பணியாற்றியவன் நான். மாநிலம் முழுதும் சுற்றுப்பயணம் மேற்கொண்டு பல்வேறு தரப்பட்ட வாசகர்களுடன் உரையாடியிருக்கிறேன். ஏராளமான புத்தகக் காட்சிகளில் பங்கெடுத்திருக்கிறேன். அந்த அனுபவம் தருகிற எண் இது. இதற்கு மேல் மிக நிச்சயமாக இல்லை. இவர்கள்தான் சமையல் நூல்கள், ஆன்மிக நூல்கள், சுய முன்னேற்ற நூல்கள், சோதிட நூல்கள், வாழ்க்கை வரலாறுகள், தொழில்சார் நூல்கள், நவீன இலக்கிய நூல்கள், கவிதைகள் எனத் தமது விருப்பத்துக்கேற்ப வாங்கி வாசிப்பவர்கள். இதில் கவனிக்க வேண்டிய முக்கியமான அம்சம் ஒன்று உண்டு. இந்த இரண்டு லட்சம் என்ற எண்ணிக்கையே மெல்ல மெல்லப் பெருகி உருவாகி வந்ததுதான். இரண்டாயிரமாவது ஆண்டுக்கு முன்னர் தமிழ் புத்தக வாசகர் உலகின் மொத்த எண்ணிக்கை ஒரு லட்சத்துக்கு மேல் கிடையாது.

நன்கு விற்கும் புத்தகம் என்றால் ஆயிரம் பிரதிகள். மிக நன்றாக விற்கும் புத்தகம் என்றால் இரண்டாயிரம் பிரதிகள். தமிழ்ப் பதிப்புலகம் எப்போதும் சொல்லும் எண்ணிக்கை இதுதான். அபூர்வமாக எப்போதேனும் ஒன்றிரண்டு புத்தகங்கள் ஐயாயிரம், ஆறாயிரம் என்ற இலக்கை எட்டியிருக்கின்றன. அது எழுதுபவரின் நட்சத்திர மதிப்பினைப் பொறுத்து நிகழ்வது. கணக்கில் கொள்ள முடியாத வகையைச் சேர்ந்தது.

இதனை இவ்வளவு உடைத்துக் காட்டுவதற்கு ஒரு காரணம் உண்டு. ஒப்பீட்டளவில் தமிழ்ச் சமூகம் வாசிப்பில் மிகவும் பின்தங்கிய சமூகமே ஆகும். புத்தகங்களின் எண்ணிக்கை இங்கே அதிகரிக்குமே தவிர, வாசக எண்ணிக்கை பெருகாது. காரணம், மிகத் தொடக்க காலம் முதலே நாம் 'பேசிக் கேட்டு'ப் பழகியவர்கள். [5] வாசித்து அறிந்து வந்தவர்கள் அல்லர். அரசியல், ஆன்மிகம் தொடங்கி அனைத்துத் துறைசார் தகவல்களையும் உரைகளின் மூலமாக, சொற்பொழிவுகளின் மூலமாகவே உள்வாங்கிப் பழகிய ஒரு மக்கள் கூட்டம், வாசிப்பு என்னும் செயலுக்குச் சுணங்குவது இயற்கை.

யுனிகோட் என்னும் புரட்சி

இந்த வழக்கம் மாறத் தொடங்கியதே இரண்டாயிரமாவது ஆண்டுக்குப் பிறகு நிகழத் தொடங்கிய டிஜிட்டல் வாசிப்புப் பழக்கத்தினால்தான். தமிழில் இது அதிவேகம் கொள்ள மூல முதற்காரணம்

யுனிகோட் என்னும் ஒருங்குறியின் வரவும் வீச்சும் என்பதில் சந்தேகமில்லை.

இணையம் இங்கே அறிமுகமான காலத்தில் தமிழில் ஒரு மின்னஞ்சல் எழுதினால் கூடவே நாம் பயன்படுத்திய எழுத்துருவை அதே அஞ்சலில் இணைத்து அனுப்பும் சூழ்நிலை இருந்தது. அஞ்சல் கிடைக்கப் பெறுபவர் அதைப் படிக்க வேண்டுமானால், நாம் இணைத்து அனுப்பிய எழுத்துருவை டவுன்லோட் செய்து, இன்ஸ்டால் செய்து ஒரு ரெஃப்ரெஷும் செய்தால்தான் சாத்தியம்.

எண்ணிப் பார்த்தால் இப்போது சிரிப்புதான் வருகிறது. ஆனால் அப்படியும் வாழ்ந்திருக்கிறோம். தமிழில் நெடுங்காலமாக அச்சிதழ் வெளியிட்டுக்கொண்டிருந்த நிறுவனங்கள் அனைத்தும் தமக்கென இணையதளம் தொடங்கியபோது ஆளுக்கொரு எழுத்துருவைப் பயன்படுத்தினார்கள். ஒவ்வொரு தளத்தைத் திறப்பதற்கும் வாசகருக்கு ஒவ்வொரு எழுத்துரு தேவைப்பட்டது.

இந்த அவலம் அனைத்தையும் யூனிகோட் துடைத்தழித்தது. தொழில்நுட்பம் அல்ல சாகசம். நுட்பத்தின் பயனை மக்கள் மத்தியில் பரவலாகக் கொண்டு சேர்ப்பதே பெருஞ்செயல். அந்த வகையில், யுனிகோடின் வரவும் பயன்பாடும்தான் தமிழ் வாசிப்பை அடுத்தக் கட்டத்துக்கு அழைத்துச் சென்றது என்று உறுதியாகச் சொல்லலாம். திண்ணை, பதிவுகள், வார்ப்பு, அம்பலம், ஊடறு, ஆறாம்திணை, தமிழோவியம் போன்ற இணையப் பத்திரிகைகள் இதன் பிறகே பெருமளவு வாசக கவனம் பெறத் தொடங்கின.

ஆனால் இரண்டாயிரத்துப் பத்தாம் ஆண்டுக்குப் பிறகு இந்த இணைய இதழ் வாசிப்பில் ஒரு தேக்கம் உருவாகத் தொடங்கியது. சமூக ஊடகங்களின் வளர்ச்சி அதன் தலையாய காரணம். கலவையான ரசனை கொண்ட அனைவரும் இணைய இதழ்களைக் காத்திருந்து வாசித்தது போக, எதுவும் நிகழும் கணத்திலேயே என்கிற புதிய சித்தாந்தம் மேலெழத் தொடங்கி, மிக விரைவில் அது அனைவரையும் கவர்ந்துகொண்டது.

தவிர, எழுதுவோர்–வாசிப்போர் என்ற இரு தரப்பாக நிகழ்ந்த ஒரு செயல்பாடு மெல்ல மெல்லத் தனது முகத்தை மாற்றிக்கொண்டு எல்லோரும் எழுதலாம், எல்லோரும் படிக்கலாம் என்கிற ஜனநாயகமயத்தின் விளைவாக மிகப் பெரிய அளவில் பிரபலமடையத் தொடங்கியது.

சமூக ஊடகங்களின் வளர்ச்சி

தமிழ்க் கணிமைச் சாதனைகளில் ஒருங்குறியைத் தொடக்கப் புள்ளியாகக் கொள்வோமானால், இந்த சமூக ஊடகப் பரவல் இன்னொரு புள்ளி. எழுதுவது என்னும் செயல்பாடு மிகச் சிலருக்கு மட்டுமே சாத்தியம், வாசிப்பது ஒன்றே வெகு மக்கள் செய்யக்கூடியது என்னும் கருத்தாக்கத்தையே தகர்த்தது இது. யாரும் எழுதலாம் என்பது மட்டுமல்ல. எதையும் எழுதலாம் என்கிற சூழலும் இதன் பின்பே உருவாகத் தொடங்கியது. அன்றாட நிகழ்ச்சிகள், சிறிய சம்பவங்கள், நினைவுக் கோவைகள், கதைகள், கவிதைகள், கட்டுரைகள், நகைச்சுவை, விமரிசனம், அரசியல், ஆன்மிகம், பொருளாதாரம், வர்த்தகம் தொடங்கி வாழ்வின் அனைத்து அம்சங்களையும் அவரவர் மொழியில் எழுதிப் பார்க்கத் தொடங்கினார்கள்.

இப்படி சமூக ஊடகங்களில் எழுத ஆரம்பித்து, இணையத்துக்கு வெளியிலும் எழுத்தாளர்களாக அறியப்பட்டவர்கள் பலருண்டு. எழுத்துத் துறைக்கு மட்டுமன்றி, இங்கிருந்து திரைத்துறைக்குச் சென்று சாதித்தவர்களும் இருக்கிறார்கள்.

அனைத்திலும் உச்சம், இன்று வெளியாகும் பெரும்பாலான வார இதழ்களில் பணியாற்றுபவர்களில் பலர் சமூக ஊடகங்களில் இருந்து கண்டெடுக்கப்பட்டவர்களே.

வலைப்பதிவில், ட்விட்டரில், ஃபேஸ்புக்கில் எழுத ஆரம்பித்த ஒருவர் தமிழின் புகழ்பெற்ற வார இதழ் ஒன்றின் பொறுப்பாசிரியராகவே ஆனார் என்பது வரலாறு.

இதுவும் ஒரு கட்டம். காட்சி ஊடகங்கள் – குறிப்பாக யூட்யூப் பிரபலமாகத் தொடங்கிய பின்பு, சமூக ஊடகங்களிலிருந்து பல பேர் அதற்குத் தாவினார்கள். பயண நேரத்தில் இரண்டு விடியோ பார்ப்பது. தூங்கப் போகும் முன் நான்கு வீடியோ பார்ப்பது. சும்மா இருக்கும் போதெல்லாம் வீடியோ பார்ப்பது.

இது பெருக ஆரம்பித்தபோது சமூக ஊடகங்களை அரசியல் கட்சிகள் குத்தகைக்கு எடுத்தன. பெரும்பாலும் அரசியல் சார்ந்த விஷயங்களே அதிகம் பேசப்பட்டன. அரசியலும் உள்ளிட்ட அனைத்தைக் குறித்தும் எழுதவும் படிக்கவும் விரும்பியவர்கள் இப்போது மீண்டும் இணைய இதழ்களைத் தேடத் தொடங்கினார்கள். இதன் விளைவாக வெளிவரத் தொடங்கியவையே சொல்வனம், தமிழினி, கனலி, நீலி, அருஞ்சொல், அகழ், மெட்ராஸ் பேப்பர் போன்ற மின்னிதழ்கள்.

வாசகர்கள் தத்தமது ரசனை சார்ந்து இதழ்களைத் தேர்ந்தெடுத்துப் படிக்கவும் இக்காலக்கட்டம் வசதியளித்தது. மறுபுறம் டெய்லி ஹண்ட் என்கிற நிறுவனம், அனைத்து அச்சிதழ்களுக்கும் டிஜிட்டல் பிரதியைத் தன்னிடம் வந்து வாசிக்க வழி செய்தது. திரள் போன்ற சில தொகுப்பு முயற்சிகள், அனைத்துச் செய்திகளையும் அவை வெளியாகும்போதே உடனுக்குடன் திரட்டி, ஒரே இடத்தில் காட்சிப்படுத்தி, வாசகரின் அலைச்சலை எளிமைப்படுத்தியது. திரள் தனது சேவைக்கு இயந்திரக் கற்றல் நுட்பத்தைப் பயன்படுத்துகிறது. அதன் மூலம் செய்திகளை வகை பிரித்து பிராந்தியவாரியாக, செய்திகளின் தன்மைவாரியாகப் பிரித்து எளிமைப்படுத்தித் தருகிறது.

வாசிப்பும் எழுத்தும் ஜனநாயகமயமானதன் நல்விளைவுகளுள் ஒன்று இது. இதன் இன்னொரு பாய்ச்சல் வேறொரு புறம் சத்தமின்றி நடந்தது. மின்நூல்கள்.

மின்நூல் வெளி

இணைய இதழ்கள் வரத் தொடங்கிய ஆரம்ப காலத்திலேயே சில மின்நூல் முயற்சிகளும் செய்து பார்க்கப்பட்டன. 'ப்ராஜக்ட் மதுரை' இதில் முதன்மையான முன்னெடுப்பு. புதிய அச்சு காணாத பண்டைய இலக்கியப் பிரதிகளைத் தேடித் தொகுத்து டிஜிட்டல் வடிவமாக இவர்கள் அளித்தார்கள். புராதனமான புத்தகங்களின்மீது ஆர்வமுள்ளோருக்கு அது பெரிய வரப்பிரசாதமாக அமைந்தது.

இரண்டாயிரத்து நான்காம் ஆண்டு தமிழோவியம் மின் இதழின் சார்பாக என்னுடைய கட்டுரைத் தொகுப்பு ஒன்றை முதல் முதலில் மின்நூலாக்கிப் பார்த்தோம். இந்தத் தொடக்க கால மின்நூல் முயற்சிகளை இப்போது எண்ணிப் பார்த்தால் திகைப்பும் வியப்புமே ஆக்கிரமிக்கின்றன. அன்று ஒரு மின்நூல் என்பது ஒரு exe file. திறந்தால் ஒரு கோப்பு வரும். உள்ளே எத்தனைக் கட்டுரைகள் அல்லது கதைகள் உண்டோ அத்தனைக்கும் தனித்தனியே ஒரு எச்.டி.எம்.எல் வடிவம் இருக்கும். பயன்படுத்தப்பட்ட எழுத்துரு தனியாக இருக்கும். அப்படியெல்லாம் உடைத்துப் பார்க்க விரும்பாவிட்டால் exe fileஐ இயக்கி நேரடியாகப் புத்தகத்தைப் படிக்கத் தொடங்கிவிடலாம். இணைய உலாவியில் அதுவும் ஒரு பக்கம் போல வந்து நிற்கும்.

எப்படியும் ஆயிரம் பிரதிகள் விற்றுவிடும் என்று எண்ணிக்கொண்டிருந்தேன். இரண்டோ மூன்றோ பிரதிகள் விற்றன என்று நினைவு. ஆனால் அந்தப் புத்தகத்தைப் பற்றிப் பலபேர் பேசினார்கள். யாஹூ குழுமங்களில் ஏராளமான மதிப்புரைகள் வெளிவந்தன. எப்படி இதெல்லாம் நடக்கிறது என்றே புரியவில்லை. பிறகு தெரிந்தது. அந்த இரண்டோ மூன்றோ நல்லவர்கள் தமது மின்னஞ்சல் பட்டியலில் உள்ள அத்தனை பேருக்கும் தாம் பெற்ற இன்பத்தைத் தள்ளி விட்டிருக்கிறார்கள்.

இது எக்காலத்திலும் எல்லாத் தளங்களிலும் இருக்கும் பிரச்னை. பைரசி. இணையத்தின் எல்லைகளற்ற வசதி வாய்ப்புகள் இத்திருட்டை இன்னும் விரிவாகச் செய்வதற்கு உதவியது. ஓசிஆர் என்னும் ஈடு இணையற்ற நுட்பம் கண்டறியப்பட்டபோது தமிழ் சமூகம் எவ்வளவு மகிழ்ச்சி கொண்டது என்பதை நாம் அறிவோம். ஆனால் அச்சுப் புத்தகங்களைப் படியெடுத்து, திருட்டுத்தனமாகச் சுற்ற விடுவதற்கே அது பெரும்பாலும் பயன்படத் தொடங்கியது.

ஒரு சம்பவம் நினைவுக்கு வருகிறது. 2005 ஆம் ஆண்டு என்னுடைய டாலர் தேசம் (அமெரிக்காவின் அரசியல் வரலாறு) புத்தகம் வெளியாகி, புத்தகக் காட்சிக்கு விற்பனைக்குச் சென்றது. ஆயிரம் பக்கப் புத்தகம். முந்நூறு ரூபாய் விலை. கண்காட்சியில் புத்தகம் நன்றாக விற்பனை ஆனது. பலரால் பேசப்பட்டது. அதுவல்ல விஷயம். கண்காட்சி முடிந்த ஒரு வாரத்தில் அந்த ஆயிரம் பக்கப் புத்தகமும் முறையாக ஒளிநகல் எடுக்கப்பட்டு அழகான பிடிஎஃப் பிரதியாக உலகெங்கும் வலம் வரத் தொடங்கிவிட்டது. இதன் உச்சம், ஒரு நண்பர் எனக்கே அந்தப் பிரதியை அனுப்பி, நன்றாக எழுதியிருக்கிறீர்கள் என்று பாராட்டவும் செய்தது.

மின்நூல்கள் வரத் தொடங்கியபோது பைரசியும் வளமாகவே வாழத் தொடங்கியது. என்னைப் போல வேறு சில எழுத்தாளர்களும் அந்நாளில் மின்நூல் வெளியிடும் முயற்சியைத் தொடங்கி, இதனாலேயே பாதியில் நிறுத்தும்படி ஆனது.

இந்தப் பிரச்னையைத் தீர்க்க ஒரே வழி, ஒவ்வொரு மின்நூல் வெளியீட்டாளரும் தத்தமது நூலை வாசிக்கத் தானே செயலியைச் சேர்த்துச் செய்து தருவதுதான் என்று முடிவு செய்தார்கள். அதாவது, குறிப்பிட்ட நிறுவனத்தின் செயலிக்குள் மட்டும்தான் அவர்கள் தரும் மின்நூலைப் படிக்க முடியும். பிரதி எடுக்க முடியாது, வினியோகம் செய்ய முடியாது.

இணையத்தில் குடிசைத் தொழில் செய்துகொண்டிருந்தோர் தொடங்கி, ஆப்பிள், கூகுள், அமேசான்வரை அனைத்துத் தரப்பினரும் இத்தகு முயற்சிகளை ஆரம்பித்தார்கள். சென்னையில் இருந்து இயங்கும் நியூ ஹொரைசன் மீடியா என்னும் நிறுவனம் நானறிந்து இப்படிப் பிரத்தியேக மின்நூல் செயலி ஒன்றைச் செய்து பார்த்தது.

ஆனால் அனைவரும் யோசிக்கத் தவறியது ஒன்றுண்டு. ஒரு வாசகன் தனது செல்போனில் எத்தனை வாசிப்புச் செயலிகளை வைத்திருக்க முடியும்? புத்தகங்கள் இடத்தை அடைத்துக்கொள்ளும்; மின்நூலில் அந்தச் சிரமம் கிடையாது என்று சொல்லிக்கொண்டு ஆரம்பித்து, மின்நூல் செயலிகளுக்கு போனில் இடம் கிடையாது என்று சொல்லும் அளவுக்கு இது போனது.

ஆனால் தீர்ப்பளிக்கும் விஷயத்தில் பயனரை விஞ்ச யாருமில்லை. எல்லா விதங்களிலும் சௌகரியமான அமேசான் கிண்டில் மின்நூல்களை ஏற்றுக்கொண்டு மற்ற அனைத்தையுமே தமிழ் வாசகர்கள் நிராகரித்துவிட்டார்கள். ஆப்பிள், கூகுள் மின்நூல்களும் இதற்குத் தப்பவில்லை என்பதே இங்கே முக்கியம்.

ஒரு விஷயம். அமேசான் கிண்டில் புத்தகங்களுக்கும் திருட்டுப் பிரதிகள் தயாரிக்க முடிந்தது. அப்படித் தயாரித்து, அவற்றை வெளியிடுவதற்கென்றே டெலிகிராமில் பல பிரத்தியேக சானல்கள் திறக்கப்பட்டன. அடையாளம் மறைத்த நபர்கள் திரை மறைவில் இருந்துகொண்டு இந்தத் திருட்டுப் பிரதிகளைத் தொடர்ந்து வெளியிட்டபோது கிண்டில் நிறுவனத்தாலும் அவர்களை ஒன்றும் செய்ய

முடியவில்லை. ஒன்றிரண்டு பிடிஎஃப் குழுக்களைப் புகார் அளித்து நீக்க முடிந்ததே தவிர, புதிது புதிதாக வேறு வேறு பெயர்களில் அவை மீண்டும் வருவதைத் தடுக்க முடியவில்லை.

அனைத்தையும் மீறி அமேசான் கிண்டில் மின்நூல்கள் மட்டும் எப்படி வெற்றி கண்டன? இதற்கு மூன்று காரணங்களைச் சொல்லலாம்.

1. பயன்பாட்டு எளிமை.

2. இயந்திரக் கற்றல் நுட்பம் மூலம் ஒரு வாசகர் ஒரு முறை தேர்வு செய்யும் புத்தகத்தைக் கொண்டு அவரது விருப்பம் அறிந்து அதற்கேற்பப் பரிந்துரைகள் செய்வது.

3. சக்தி மிக்க தேடுபொறி வசதி.

கிண்டிலின் வரவு, தமிழ் வாசிப்பு வரலாற்றில் சந்தேகமின்றி, ஒரு முக்கியமான புள்ளி. குறிப்பாக, அவர்கள் தருகிற வாடகை நூலக வசதி. அமேசான் நிறுவனம் ஆண்டுதோறும் நடத்தும் நாவல் போட்டிகள் குறித்து அறிவீர்கள். ஓராண்டு அந்தப் போட்டிக்கு நடுவராக இருக்கும் வாய்ப்பு எனக்குக் கிடைத்தது. அப்போது கிண்டில் தமிழ்ப் பிரிவின் உயரதிகாரிகளுடனும் தொழில்நுட்ப வல்லுநர்களுடனும் கலந்து பேசி அதன் செயல்பாட்டினை ஓரளவு விளங்கிக்கொள்ள முடிந்தது.

அதிகம் படிக்காத, எளிய வேலைகளுக்கு மணிக்கணக்கில் பேருந்து அல்லது ரயில் பயணம் செய்து திரும்பும் பெண்களே கிண்டில் வாடகை நூலகத்தின் பெரும்பான்மை வாசகர்களாக இருக்கிறார்கள். கையில் ஒரு போனும் மாதம் நூற்றைம்பது ரூபாய் சந்தா தொகையும் இருந்தால் போதும், எவ்வளவு வேண்டுமானாலும் படிக்கலாம் என்பதை அவர்கள் தமக்குக் கிடைத்த வரமாகப் பார்க்கிறார்கள். முன்னொரு காலத்தில் தமிழ் வார இதழ்களில் கோலோச்சிய பெண் எழுத்தாளர்களை அடியொற்றி, இந்தப் புதிய தலைமுறை வாசகர்களுக்காகக் கதைகள் எழுதவென்றே நூற்றுக் கணக்கான புதிய பெண் எழுத்தாளர்கள் அங்கேயே பிறந்து வளர்ந்திருக்கிறார்கள். எளிய குடும்பக் கதைகள். எளிய காதல் கதைகள். இவற்றைத் தவிர வேறெதுவும் இல்லை. பெரிய மொழி அறிவோ, இலக்கண அறிவோ, புனைவாற்றலோ இந்த எழுத்தாளர்களுக்குக் கிடையாது. ஆனால் கதைகளை வாழ்க்கையில் இருந்து எடுக்க வேண்டும் என்கிற சூட்சுமம் மட்டும் தெரியும். கிண்டிலில் வாழும் ஏராளமான பெண் எழுத்தாளர்கள் நூற்றுக் கணக்கான கதைகளை (அவர்கள் நாவல் என்பார்கள்) எழுதி வெளியிட்டிருக்கிறார்கள். கிண்டில் வாசகர்களிடையே அவர்கள் பெருநட்சத்திரங்கள். ஆனால் மின்நூல் உலகுக்கு வெளியே வசிக்கும் யாருக்கும் அந்த எழுத்தாளர்களின் பெயர்கள்கூடத் தெரிந்திருக்க வாய்ப்பில்லை.

செயற்கை நுண்ணறிவு என்னும் சாகசம்

பெயரில் என்ன இருக்கிறது? அல்லது பெயரேதான் எதற்கு? சாட் ஜிபிடியின் வரவுக்குப் பிறகு என்ன கேட்டாலும் சில வினாடிகளில் கிடைத்துவிடும் என்றாகிவிட்டது. உள்ளே உட்கார்ந்துகொண்டு எழுதுபவர் யார்? தெரியாது. அவருக்கு எப்படி உலகில் உள்ள எல்லாவற்றைப் பற்றியும் ஏதோ கொஞ்சமாவது தெரிந்திருக்கிறது? தெரியாது. கேள்வி கேட்டால் பதில் சொல்கிறது. கதை எழுதச் சொன்னால் எழுதுகிறது. கட்டுரை கேட்டால் தருகிறது. கவிதை வேண்டுமானாலும் உடனே தயார். முன்னும் பின்னும் சிறிது தட்டி சரி செய்ய வேண்டிய வேலை மட்டும்தான் நமக்கென மிச்சம் இருக்கிறது. ஏழாம் அறிவு என்றே சொல்லப்படும் செயற்கை நுண்ணறிவுத் தொழில்நுட்பம் நம்மை இப்போது ஆளத் தொடங்கியிருக்கிறது. நான் எழுதுவதை நீ படி என்ற காலம் இனி இல்லை. உனக்கு என்ன பிடிக்கும் அல்லது இந்தக் கணம் என்ன படிக்க விரும்புகிறாய் என்று சொல், அதனைத் தருகிறேன் என்று கேட்டு உபசரிக்கக் கூடிய காலம் மிகச் சமீபத்திலேயே உள்ளது.

தமிழைப் பொறுத்தவரை இந்த நுட்பம் இன்னும் அவ்வளவு துல்லியமாகவில்லை என்றாலும் அது நிகழக்கூடிய காலம் குறைவானதாகவே இருக்கும். ஆழி செந்தில்நாதனின் 'ஐலேசா' [6] போன்ற முயற்சிகள் இந்நம்பிக்கையை வலுவாகவே அளிக்கின்றன. செயற்கை நுண்ணறிவு நுட்பத்தைப் பயன்படுத்தி, தேவையான எந்தத் துறை சார்ந்தும் புதிய பிரதிகளை உருவாக்க முடியும் என்கிறது ஐலேசா. இயந்திரம் கற்கும் என்றால் இயந்திரம் படைக்கவும் செய்யவல்லதுதானே? 'கண்டெண்ட் க்ரியேஷன்' என்பதை இயந்திரத்தின் பொறுப்பில் விட்டுவிட்டு, அதனைச் சரி செய்து, சுவாரசியமாக்கும் பணியை மட்டும் எழுத்தாளர்கள் எடுத்துக்கொண்டால் போதும் என்றொரு காலம் வந்தேவிடும் என்று தோன்றுகிறது. இதற்கு இன்னோர் உதாரணமும் சொல்லலாம்.

முன்னொரு காலத்தில் தேவநாராயணன் என்றொரு எழுத்தாளர் இருந்தார். சிறந்த கவிஞரும் கூட. தமிழ் சினிமா துறையில் அவரை அறியாதவர்கள் இருக்க முடியாது. இங்கே வெளியாகும் பெரும்பாலான டப்பிங் படங்களுக்கும் அவர்தான் வசனம் எழுதுவார். அது டப்பிங் படம் என்றே தோன்றாத அளவுக்குத் தரமாக அது இருக்கும். அவரைப் பின்பற்றி எவ்வளவோ டப்பிங் வசனகர்த்தாக்கள் திரைத் துறைக்கு வந்து வாழ்ந்துவிட்டுச் சென்றிருக்கிறார்கள்.

ஆனால் இன்றைக்கு மொழி மாற்றம் என்பது ஒரு கணினிச் செயல்பாடு மட்டுமே. அதற்கு மொழியறிந்த எழுத்தாளர் ஒருவர் தேவையில்லை என்ற நிலை உருவாகிவிட்டது. நெட் ஃப்ளிக்ஸ், அமேசான் போன்ற

பல பன்னாட்டு நிறுவனங்களே படங்களுக்கான சப் டைட்டில்களுக்கு கூகுள் மொழிபெயர்ப்பைத்தான் பயன்படுத்துகின்றன.

இது வேறு உலகம்

நவீன வாழ்க்கை நமது ரசனைகளை விரிவாக்கியிருக்கிறது. ரசனை விரிவடையும்போது தேவைகளும் விரிகின்றன. அவசர யுகத்தில் நாம் வாழ்ந்துகொண்டிருக்கிறோம். எதையும் நிறுத்தி, நிதானமாக என்பதே பெரும்பாலும் வழக்கொழிந்துகொண்டிருக்கிறது. பெண்கள் வீட்டில் இருந்து வீட்டை கவனித்துக்கொள்வார்கள், ஆண்கள் சம்பாதிக்கச் செல்வார்கள் என்பதே ஒரு மாய யதார்த்தம் போலத் தோன்ற ஆரம்பித்துவிட்டதல்லவா?

காலை நடைப்பயிற்சிக்கு வருவோரை எதிர்பார்த்துக் காய்கறிக்கடைக்காரர் கடை விரிக்கிறார். நறுக்கிய கேரட், பீன்ஸ், வெண்டைக்காய்த் துண்டுகள் தனித்தனி பாக்கெட். நறுக்கிய வெங்காயம் ஒரு பாக்கெட். தோல் உரித்த பூண்டு ஒரு பாக்கெட். கீரை என்றால் ஆய்ந்து, நறுக்கி, அப்படியே கொதிக்கும் நீரில் போட வசதியாகத் தனியே ஒரு பாக்கெட்.

சமைக்கவும் நேரமில்லாவிட்டாலும் பிரச்னை இல்லை. செயலி இருக்கிறது. அழைத்து ஆணையிட்டால் ஐந்து பத்து நிமிடங்களில் சிற்றுண்டியோ, பேருண்டியோ வாசலுக்கு வந்துவிடுகிறது.

பேருந்து நிலையம் சென்று காத்திருக்க வேண்டாம். ஆட்டோ வருகிறதா என்று வீதி பார்த்திருக்க வேண்டாம். டாக்டர் இருப்பாரா என்று விசாரிக்க வேண்டாம். எல்லாம் செயலிக்குள் இருக்கிறது. கையில் செல்போன் இருந்து, செல்போனில் வேண்டிய செயலிகள் இருந்தால் முடிந்தது வேலை.

வாசிப்பும் இப்படித்தானே உருமாற்றம் கொள்ள வேண்டும்? தமிழைப் பொறுத்தவரை வாசிப்பவர்கள் எக்காலத்திலும் சிறுபான்மையினரே என்று முன்னர் சொன்னேன். அந்தச் சிறுபான்மை சமூகம் தனது வாசிப்பை இன்றைக்கு இப்படி நவீனமாக்கிக்கொண்டிருக்கிறதே தவிர, வாசிக்காமல் இல்லை. இணைய இதழ்களிலும் சமூக ஊடகங்களிலும் மின்நூல்களிலும் வாசக கவனம் திரும்பியதன் முதன்மைக் காரணம், அவற்றின் பயன்பாட்டு எளிமை. மொபைல் இல்லாமல் இன்று யாருமில்லை. எனவே மொபைலுக்குள் கிடைக்கும் எதுவும் தத்தமது வாசகரை / வாடிக்கையாளரைச் சென்று சேரும் என்பதே உண்மை.

இதுதான் இனி நிரந்தரமா என்றால், அப்படிச் சொல்லிவிட முடியாது. இது ஒரு பாய்ச்சல். அசுரப் பாய்ச்சல் என்பதில் சந்தேகமில்லை. ஆனால் தொழில்நுட்பம் இன்னும் இன்னும் புதிய எல்லைகளைத் தேடிப் பறந்துகொண்டே இருக்கிறது. ஒரு மொபைல் போனைக் காட்டிலும் அத்தியாவசியம் என்று கருதக்கூடிய ஒரு கருவி நாளை வரலாம். இன்றைய நுட்பங்களின் 2.0 அதன் அணிகலனாகலாம்.

யார் கண்டது? இப்போது நீங்கள் பொருட்படுத்தி வாசிக்கும் இந்தக் கட்டுரை அப்போது எழுத்தாகவும் அல்லாமல் ஒலியாகவும் அல்லாமல் வெறும் உணர்வாகத தானே மிதந்து வந்து சத்தமே இல்லாமல் உங்கள் சிந்தனைக்குள் அமர்ந்துகொள்ளலாம். மனத்துக்குள்ளேயே நீங்கள் ரசித்துக் கைதட்டினால் அதுவும் சத்தமேயின்றி என்னை வந்தடைந்து புல்லரிக்கச் செய்யலாம்.

எல்லையற்ற சாத்தியங்களின் காலத்தில் வாழ்கிறோம். எதையும் புரிந்துகொண்டு ஏற்பதே அடுத்தக்கட்ட வளர்ச்சிக்கு அடித்தளம்.

நிறைவாக

இந்தக் கட்டுரையின் தொடக்கத்தில் தமிழின் முதல் அச்சுப் புத்தகத்தைப் பற்றிக் குறிப்பிட்டேன். லத்தீன் லிபியில் எழுதப்பட்ட தமிழ்ப் புத்தகம். எல்லா புராதனமான நடைமுறைகளும் நாகரிக உலகில் மறு ஆக்கம் பெற்றுக்கொண்டிருக்கும் காலம் இது. நீங்கள் கவனித்துப் பார்க்கலாம். லட்சக் கணக்கான, கோடிக்கணக்கான இளையதலைமுறையினர் – இதில் பால் பேதமே இல்லை – தமிழ்தான் பேசுகிறார்கள். தமிழில்தான் உரையாடுகிறார்கள். அவர்களது வாட்சப், மெசஞ்சர் உள்ளிட்ட எந்த சாதனத்தைத் திறந்து காட்டச் சொல்லிப் பார்த்தாலும் நமக்கு இது விளங்கிவிடும். தமிழ்தான். ஆனால் ஆங்கில லிபியில் எழுதப்படும் தமிழ்!

இது இத்தலைமுறையின் பிரச்னை அல்ல. நமது கல்வி முறையின் சிக்கல். ப்ரீ கேஜி வயதிலிருந்தே ஆங்கில வழிக் கல்வி அவர்களுக்கு வழங்கப்படுகிறது. எதிர்கால நலனை உத்தேசித்து இரண்டாவது மொழியாக பிரெஞ்சு அல்லது ஜெர்மனைத் தேர்ந்தெடுக்கிறார்கள். மூன்றாவதாகவும் ஒரு மொழி என்றால் ஹிந்தியைத் தேர்ந்தெடுக்கிறார்கள். இதனால் வீட்டில் பேசப்படும் மொழியே ஆனாலும் அதன் எழுத்து வடிவம் மாணவர்களுக்கு அறிமுகமில்லாமல் போகின்றது. தமிழ் பிடிக்காமல் அவர்கள் தமிழில் எழுதாமல் இல்லை. தமிழ் அவர்களுக்குத் தரப்படுவதில்லை; அதனால் அவர்கள் பயன்படுத்துவதில்லை.

இன்றைய சூழலில் தமிழ்நாட்டில் அரசுப் பள்ளிகளில் படிக்கும் மாணவர்கள் மட்டும்தான் ஓரளவேனும் தமிழ் வாசிக்கக் கூடியவர்களாக இருக்கிறார்கள். அவர்களும் மேற்படிப்பு என்று போகும்போது ஆங்கில வழிக் கல்வியையே தேர்ந்தெடுத்துக்கொள்கிறார்கள். இது தவிர்க்க முடியாதது. தமிழில் வாசிப்பு குறைந்து வருவதாகச் சொல்லப்படுவதன் அடிப்படைக் காரணம் இதுதான்.

ஒரு மொழியின் செழிப்பும் வளர்ச்சியும் அதில் படைக்கப்படும் இலக்கியங்களால் தீர்மானிக்கப்படுவதாகச் சொல்வார்கள். உணர்ச்சிவசப்படாமல் யோசிக்க முடியுமானால், ஒரு மொழியின் இருப்பும் செழிப்பும் வளர்ச்சியும் அதைப் பேசியும் எழுதியும் படித்தும் வாழக்கூடிய மக்களால் தீர்மானிக்கப்படுவதுதான். நமக்கு நன்கு தெரிந்த, இலக்கிய வளம் மிக்க சமஸ்கிருதத்தை முன்வைத்தே இதனைப் புரிந்துகொள்ள முடியும். எந்த இலக்கியமும் அதனைத் தலைமுறை தோறும் வாசிப்பதற்கு வருகிற வாசகராலேயே தழைக்கின்றன.

ஆனால், தமிழை ஆங்கில லிபியில் பயன்படுத்தும் தலைமுறை எப்படி வாசிப்புக்குத் திரும்பும்?

என்றால், எல்லாம் வல்ல ஏஐ இவ்விஷயத்தில் எதிர்காலத்தில் உதவி செய்யக்கூடும். இளைய தலைமுறையினர் – இன்றைய தலைமுறையினர் படித்து ரசிக்க விரும்பக்கூடிய வடிவத்தில் அது உருமாற்றித் தரும்.

எண்ணிப் பார்த்தால் வியப்பாகத்தான் இருக்கிறது. செயற்கை நுண்ணறிவு ஒரு பிரதியைத் தானே உருவாக்கும். அதை எம்மொழிக்கு வேண்டுமானாலும் மாற்றும். எந்த வடிவத்தில் வேண்டுமானாலும் எடுத்துத் தரும். எல்லாமே சில மணித்துளிகளில் நிகழக்கூடியவையாக இருக்கும்.

எத்தனை பெரிய முன்னேற்றம் இது. Content ஆளும் உலகம் இது. எல்லா துறைகள் சார்ந்தும் எல்லா விதமான தேவைகளுக்கேற்பவும் எந்த வடிவத்திலும் இது பிரதிகளை உருவாக்கித் தரும் என்பது பத்திரிகை, பதிப்புச் சூழலுக்கு மிகவும் சாதகமான அம்சம்.

ஏனெனில் கலைச் செல்வங்கள் யாவும் கொணர்ந்திங்கு சேர்க்க பாரதியின் காலத்தைப் போல இனி சிரமம் இல்லை. ஒரு கமாண்ட் போதும். செயற்கை நுண்ணறிவுத் தொழில்நுட்பத்தைக் குறித்து, அதனிடமே ஒரு கட்டுரை எழுதச் சொல்லி, அது சரியாக எழுதியிருக்கிறதா என்று நாம் பரிசீலனை செய்தால் போதும். பதிப்பு, பத்திரிகைத் துறைகள் மட்டுமல்ல. காட்சி ஊடகங்கள், வானொலி, தொலைக்காட்சி உள்பட எங்கெல்லாம் 'கண்டெண்ட்' முக்கியம் என்று கருதப்படுகிறதோ, அங்கெல்லாம் ஒரு கையாளாக இந்த நுட்பம் இனி பயன்படும்.

தமிழ் வாசகர்களைப் பொறுத்த அளவில், வாசிக்கத் தெரிந்த தலைமுறையினருக்குப் பிரச்னையே இல்லை. இரண்டாயிரமாவது ஆண்டுக்குப் பிறகு உருவாகி வந்த தலைமுறைக்கு, அவர்கள் வாசிக்க விரும்பும் வடிவத்தில் எதையும் உருமாற்றித் தரும் பொறுப்பை அநேகமாக AI இனி எடுத்துக்கொள்ளும் என்று நினைக்கிறேன். இன்றைக்கு கூகுளின் தமிழாக்கக் குளறுபடிகளை நாம் கிண்டல் செய்துகொண்டிருக்கிறோம். இது ஒரு தொடக்கக்கட்ட முயற்சி, ஆனால் மிகப்பெரிய நல்விளைவுகளை எதிர்காலத்தில் கொண்டுவரக்கூடியது என்பதை எண்ணிப் பார்க்க மறந்துவிடுகிறோம்.

இதே போல நாளை ஏஐ உருவாக்கும் 'தங்கிலீஷ்' பிரதிகளையும் கிண்டல் செய்வோம். ஆனால் நிதானமாக, மௌனமாக அது தமிழில் வாசிப்போர் எண்ணிக்கையை அதிகப்படுத்தி, வாசிப்பில் ஆர்வத்தை உண்டாக்கி, முறையாகத் தமிழ் பயிலவும் காரணமாக விளங்கப் போகிறது என்பதில் சந்தேகமில்லை.

முன் சொன்னதுதான். இரண்டாயிரத்துக்குப் பிறகுதான் தமிழில் வாசிப்போர் எண்ணிக்கை அதிகரித்திருக்கிறது. இன்னும் இருபதாண்டுகளில் இது இன்னும் பல மடங்காகுமே தவிர குறையாது. ஆனால் எத்தனை ஆயிரம் பக்கங்களானாலும் அது ஒரு கைபேசிக்குள் அடங்கும் நுட்பத்தைத் தாங்கியதாக இருக்க வேண்டும்.

காலம் இதனைத்தான் எதிர்பார்க்கிறது. நுட்பம் இதை நோக்கித்தான் நகர்ந்துகொண்டிருக்கிறது.

தொடர்புடைய சுட்டிகள்:

1. தமிழ் அச்சிடல் வரலாறு தகவல் தளம் https://amarkkalam.forumta.net/t4940-topic
2. தினத்தந்தி https://en.wikipedia.org/wiki/Dina_Thanthi
3. தமிழ் பேசுவோர் World data https://www.worlddata.info/languages/tamil.php#:~:text=The%20Tamil%20language%20(native%20name,Tamil%20as%20their%20mother%20tongue.
4. all about book publishing https://www.allaboutbookpublishing.com/10309/chennai-international-book-fair-2023-a-new-chapter-in-tamil-publishing/#:~:text=In%202021%2C%20around%2017%2C000%20books,other%20Indian%20and%20international%20languages.
5. An indicator will be 26M subscribers for Sun Music Channel in YouTube. https://socialblade.com/youtube/channel/UCBnxEdpoZwstJqC1yZpOjRA
6. ஐலசா https://www.ailaysa.com

TAMIL E-LEARNING PLATFORMS AND TOOLS

An Immersive Journey: Tamil Epic Poetry Silapathikaram Stepping into Metaverse

R. Rajkumar, Dominic Dunn, Antony Sam Jaiton

ABSTRACT:

Silapathikaram, the renowned Tamil epic poem, stands as a testament to the vibrant culture and rich literary heritage of India. In a world increasingly dominated by digital technologies, preserving and transmitting this legacy to new generations poses a significant challenge. This research proposes a novel approach utilizing the immersive potential of the metaverse to create an interactive role-playing experience that transports learners into the captivating world of Silapathikaram. The objectives are to develop a captivating metaverse environment that recreates the vivid scenes, characters, and events of Silapathikaram with a high degree of historical and cultural accuracy. To design interactive role-playing scenarios that allow learners to engage with the narrative, embodying characters, making choices, and experiencing the consequences of their actions.

The methodology for Metaverse development includes creating a detailed virtual world encompassing the key settings of Silapathikaram, incorporating architectural styles, landscapes, and cultural artifacts from the Chola period. Character creations are to develop avatars that learners can personalize, allowing them to embody various characters from the epic, such as Kannagi, Kovalan, Madhavi, and chola king Karikalan. A thriving online community of learners and educators dedicated to preserving and promoting the rich heritage of Silapathikaram. A pioneering model for utilizing the metaverse as a tool for cultural education and heritage preservation, with potential applications for other historical and literary works. Stepping into Silapathikaram through the metaverse presents a unique opportunity to bridge the gap between tradition and modernity, fostering a deeper understanding and appreciation for Tamil culture and literature among new generations. By harnessing the power of virtual reality and interactive storytelling, this project has the potential to revolutionize cultural education and inspire a renewed interest in the timeless masterpiece of Silapathikaram.

R. Rajkumar, Assistant Professor, Department of DSBS, School of Computing, SRMIST, India. Email: rajkumar2@srmist.edu.in
Dominic Dunn, Principal Lecturer (International), Department of Digital Arts and Animation, Centre for Digital Innovation, Teesside University, Middlesbrough, UK. Email: dominic.dunn@tees.ac.uk
Antony Sam Jaiton, B.Tech in Gaming Technology, Department of DSBS, School of Computing, SRMIST, India. Email: ap3723@srmist.edu.in

1. INTRODUCTION

The virtual world should be a meticulously crafted learning experience, one that seamlessly blends cutting-edge technology with the timeless wisdom of the epic. Across millennia, the human spirit has craved tales woven from threads of history, myth, and imagination. These narratives, passed down through generations, pulsate with the collective memory of a people, their triumphs and tragedies etched in verse and rhyme.

Among these luminous tapestries of storytelling, few shine as brightly as Silapathikaram, the Tamil epic poem that has dazzled readers for centuries. Now, prepare to transcend the limitations of the printed page, for Silapathikaram is poised to take a giant leap into the future, stepping into the captivating realm of the metaverse. Imagine a world where we don't merely read about Kannagi's fiery spirit or Kovalan's ill-fated journey; to inhabit them. Where the bustling streets of Madurai unfold before you, fragrant with spice and alive with the clamor of trade. Where Karikalan's opulent palace rises in all its majestic splendor, its walls whispering tales of ancient kings and forgotten battles. This is the promise of Silapathikaram in the metaverse – an immersive odyssey into the beating heart of Tamil culture, where history and fiction intertwine in a breathtaking tapestry of virtual reality.

The potential for gaming and entertainment is undeniable, spatial computing's true power lies in its ability to transform various industries:

1.1 Spatial computing

It is still in its early stages, but the pace of development is rapid. Advancements in hardware, software, and sensor technology are pushing the boundaries of what's possible, and major tech giants are pouring billions into making this technology accessible to everyone.

Education domain: The students exploring the pyramids of Egypt on a virtual field trip or dissecting a virtual frog in biology class. Spatial computing can bring abstract concepts to life, making learning more interactive and engaging.

Architecture and Design: Architects can walk through their creations before they're even

built, and designers can prototype products in real-time using AR.

Manufacturing and Engineering: Workers can receive real-time instructions and guidance through AR overlays, and engineers can collaborate on complex projects from anywhere in the world using VR.

2. METHODOLOGY

Craft your own personalized avatar, embodying iconic characters like Madhavi, the enigmatic dancer, or the cunning courtier, Madalan. Walk alongside Kannagi as she seeks justice, navigate the treacherous seas with Kovalan, or engage in spirited debates with the scholars of the Chola court. Each step you take, each choice you make, becomes a thread woven into the intricate narrative tapestry.

The implementation steps for designing a Silapathikaram metaverse drama using Unity:

2.1 Planning and Research:

Deep Dive into Silapathikaram: Thoroughly study the epic's narrative, characters, themes, and cultural context.

Story Segmentation: Identify key scenes, plot points, and character interactions that translate well into interactive drama.

Audience Definition: Determine the target audience and tailor the experience accordingly (e.g., educational, cultural immersion, entertainment).

Technical Considerations: Research hardware requirements, VR platforms, and Unity features for metaverse development.

2.2 Worldbuilding and Environment Design:

Virtual Landscapes: Create immersive, visually stunning environments that capture the essence of the Chola dynasty, including:

- Madurai city streets
- Karikalan's palace
- Lush forests and natural landscapes

Architectural Accuracy: Research and incorporate architectural styles, patterns, and materials specific to the Chola period.

Environmental Storytelling: Infuse the environments with subtle details that reinforce the narrative and historical context.

2.3 Character Design and Animation:

3D Modeling: Craft highly detailed 3D models of Silapathikaram's characters, including Kannagi, Kovalan, Madhavi, Karikalan, and other significant figures.

Rigging and Animation: Implement lifelike movements and expressions through rigging and animation techniques.

Character Customization: Consider allowing users to create personalized avatars or choose from a range of pre-designed characters.

2.4 Interactive Storytelling and Quest Design:

Branching Narratives: Allow users to influence the story's direction through choices and actions.

Quests and Challenges: Integrate engaging quests that encourage exploration, problem-solving, and interaction with characters and environments.

Dialogue Systems: Develop natural and meaningful conversations with characters through interactive dialogue trees.

2.5 Audio Design and Soundscape

Immersive Audio: Create a captivating soundscape that transports users to the world of Silapathikaram, including:

- Ambient sounds of nature and city life
- Character voices and dialogue
- Traditional Tamil music and instruments

Spatial Audio: Utilize spatial audio techniques to enhance immersion and create a sense of presence in the virtual world.

2.6 User Interaction and Control

VR Integration: Implement seamless integration with VR headsets and controllers for a fully immersive experience.

Intuitive Controls: Design intuitive controls for movement, interaction, and inventory management.

Multiplayer Functionality: Consider enabling multiplayer interactions for collaborative storytelling and social experiences.

2.7 Testing and Refinement

Rigorous Testing: Conduct comprehensive testing with diverse users to identify and address any bugs, glitches, or usability issues.

User Feedback: Gather feedback from target audiences to refine the experience and ensure its cultural authenticity and engagement.

Iteration and Improvement: Continuously iterate on the design and implementation based on feedback and testing results.

2.8 Customized information

Cultural Sensitivity: Respectfully represent Tamil culture and history, consulting with experts for guidance.

Educational Resources: Incorporate educational elements to enhance understanding of Silapathikaram and its significance.

Accessibility: Design for inclusivity, considering users with disabilities and varying levels of technical expertise.

Cross-Platform Compatibility: Explore options for making the experience accessible across different VR platforms and devices.

3. IMPLEMENTATION

Stepping into the metaverse through Silapathikaram is more than just entering a virtual world; it's a transformative act. It's about bridging the gap between past and present, ensuring that the timeless wisdom of ancient epics continues to illuminate our path forward. It's about fostering a deeper understanding and appreciation for Tamil culture and literature, not as relics of the past, but as vibrant forces shaping the future. It's about igniting a renewed passion for learning, where technology becomes a bridge, not a barrier, to the wellsprings of knowledge. This journey doesn't end with the individual. The metaverse fosters a vibrant online community, a digital Madurai where learners and educators from across the globe gather to celebrate and explore the rich heritage of Silapathikaram. Imagine lively forums buzzing with interpretations and insights, virtual classrooms echoing with scholarly discourse, and collaborative projects that breathe new life into the epic's enduring themes. This community becomes a crucible where tradition and modernity forge a powerful alliance, ensuring that the timeless wisdom of Silapathikaram resonates with new generations.

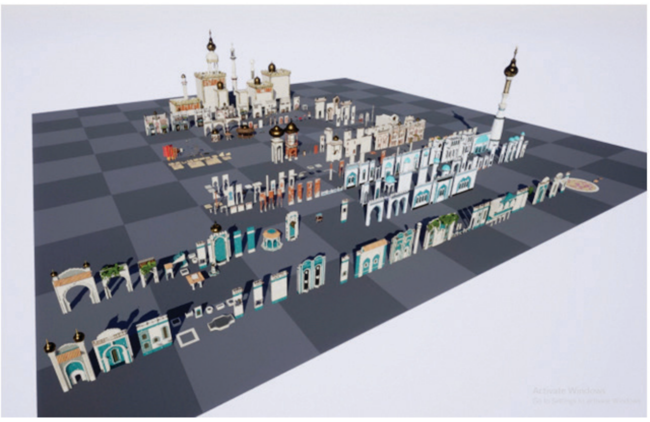


Fig 1.1: Overall Architecture

The heart of the metaverse, the world engine, manages the simulation of the virtual environment. It powers:

Physics and Rendering: Realistic physics engine simulates gravity, collisions, and object interactions. Efficient rendering engine generates high-quality visuals, lighting, and shadows.

Spatial Audio: Creates immersive soundscapes with accurate positioning and dynamic changes based on movement and environment.



Fig. 1.2: Realistic Physics

3.1. Content Servers:

These servers store and deliver various assets that populate the metaverse, including:

3D Models: Characters, environments, objects, and props are meticulously crafted 3D models optimized for performance.

Textures and Materials: Realistic textures and materials add depth and detail to the virtual world.

Animations and VFX: Character animations, environmental effects, and combat sequences bring the world to life.



Fig 1.3: 3D Model of Kovalan

Audio Assets: Voice acting, ambient sounds, and music create a captivating soundscape.

Script Logic and AI: Scripted events, character behaviors, and AI routines drive the narrative and create interactive experiences.

3.2 Administration and Management Tools:

This layer provides backend tools for managing the metaverse:

User Management: Create and manage user accounts, track progress, and personalize experiences.

Content Management: Update and add new content, manage assets, and maintain world consistency.

Monitoring and Analytics: Track user behavior, performance metrics, and identify areas for improvement.

Security and Moderation: Ensure a safe and secure environment for all users, implement moderation tools to prevent unwanted behavior.

3.3 Community and Social Features:

To foster a vibrant community, the metaverse might integrate features like:

Multiplayer Interactions: Users can collaborate on quests, engage in roleplay, and socialize with other players.

Voice Chat and Communication: Real-time voice chat and text communication options enhance social interaction and collaboration.

Leaderboards and Achievements: Encourage healthy competition and track progress through leaderboards and achievement systems.

Community Events and Activities: Regular events, quizzes, and challenges keep the community engaged and offer new experiences.

4. GENERATIVE AI IN CREATION

has the potential to revolutionize the gaming industry by introducing a level of dynamism and personalization never before seen. Here are some ways it can be utilized:

4.1 Content Creation:

World building: Generate vibrant and endlessly diverse landscapes, dungeons, and cities, creating unique experiences for each player.

Character creation: Design personalized characters with unique appearances, back stories, and abilities, fostering deeper player connection.

Storytelling: Craft dynamic narratives that adapt to player choices and actions, making every play through feel different.

Quest generation: Create custom quests tailored to player preferences and level, ensuring fresh challenges and engagement.

Procedural generation: Generate unique items, weapons, and enemies on the fly, adding an element of surprise and discovery.

4.2 Enhanced Gameplay:

AI-powered NPCs: Develop intelligent and adaptable non-player characters that react realistically to player choices and situations.

Real-time world reactions: Create dynamic environments that change based on player actions, offering consequences and rewarding exploration.

Personalized difficulty: Adjust gameplay difficulty levels dynamically based on player skill and preferences.

Improved balance: Use AI to analyze gameplay data and automatically balance game mechanics for a fair and engaging experience.

4.3 Player Experience:

Adaptive music and sound design: Generate dynamic soundtracks and sound effects that respond to player actions and the environment, creating a more immersive experience.

Procedural voice acting: Create real-time voice narration and dialogue that reacts to player choices and adds to the immersion.

Personalized tutorials and guidance: Use AI to adapt tutorials and in-game help to individual player needs and learning styles.

Virtual companions: Develop AI-powered companions that accompany players on their journey, offering assistance, advice, and even emotional support.

5. CONCLUSION

This project is not merely a digital rendering of an ancient epic; it's a portal to a lost world, a vibrant tapestry woven from history, technology, and imagination. Within the metaverse, Silapathikaram transcends the confines of text, offering a profound and immersive experience unlike any other.

By marrying meticulous historical accuracy with cutting-edge virtual reality, this venture unlocks a treasure trove of knowledge, inviting learners to explore the bustling streets of Madurai, delve into the opulent palace of Karikalan, and stand alongside iconic characters like Kannagi and Kovalan. It's an education not passive, but dynamic, where history is breathed, not merely read. In the metaverse, the epic transcends mere entertainment; it becomes a living legacy, an inspiration, and a testament to the enduring power of human storytelling.This project marks a pivotal moment, not just for Tamil culture, but for the future of cultural education itself. It demonstrates the unparalleled potential of the metaverse to bridge the gap between tradition and modernity, rekindling interest in timeless masterpieces for new generations.

REFERENCES

1. Trends in Research and Application (2023) by Michael Jones, et al. in IEEE Transactions on Games
2. Building a Shared Virtual World for Collaboration and Play (2022) by Sarah Smith, et al. in IEEE Computer
3. Enabling Immersive Experiences for Education and Entertainment (2022) by David Lee, et al. in IEEE Transactions on Emerging Topics in Computing
4. Challenges and Opportunities for Building a Persistent Immersive World (2023) by Daniel Brown, et al. in IEEE Access
5. Engaging Users with Interactive Experiences (2023) by Emily Garcia, et al. in IEEE Transactions on Learning Technologies
6. Leveraging VR and AR for Immersive Narratives (2022) by Christopher Davis, et al. in IEEE Transactions on Visualization and Computer Graphics
7. The Future of Competitive Gaming (2023) by Matthew Williams, et al. in IEEE Transactions on Games
8. Educational Design Principles for the Metaverse (2022) by Jennifer Hernandez, et al. in IEEE Transactions on Education
9. Protecting Users in a Virtual World (2023) by Thomas Miller, et al. in IEEE Transactions on Human-Machine Systems
10. Creating a Multi-sensory Experience for All (2022) by Katherine Johnson, et al. in IEEE Access
11. Challenges and Opportunities (2023) by Elizabeth Moore, et al. in IEEE Transactions on Technology and Society
12. Designing Responsible Gameplay in the Metaverse (2022) by William Robinson, et al. in IEEE Transactions on Games

Dynamic Language Learning: Immersive Tamil Education through 3D Visualization in English-Tamil Flashcards App

Yuvasree P, Kavi Priya B, Thenmozhi K

ABSTRACT

The main objective of this application is to introduce an interactive flashcard designed for learning English-Tamil vocabulary. Utilizing the Tkinter library for the graphical user interface, the app dynamically generates flashcards with random English words to their corresponding Tamil translations. Users can input their word pairs, and the application supports audio pronunciation playback for both English and Tamil translations. Additionally, a 3D viewer feature is included, offering a visually engaging way to display word images. The integration of text-to-speech functionality enhances the language learning experience, making the application an interactive and educational tool.

Yuvasree P, Kavi Priya B, Thenmozhi K
UG Student, Department of Artificial Intelligence and Data Science,
Dr. N.G.P Institute of Technology, Coimbatore.
{22ad062, 22ad023, 22ad055}@drngpit.ac.in

1. INTRODUCTION

The "English-Tamil Flashcards" application represents an innovative educational tool with a singular aim – to revolutionize language learning. Through the integration of the Tkinter library, this interactive platform dynamically generates flashcards, presenting users with random English words and their corresponding Tamil translations. Notable is its inclusivity, allowing users to contribute their word pairs and shaping a personalized learning journey. The application's commitment to auditory learning shines through with its support for audio pronunciation playback for both English and Tamil translations, providing a comprehensive language immersion. An innovative 3D viewer feature takes the educational experience to new heights, offering a visually stimulating showcase of word images. Meanwhile, the integration of text-to-speech functionality further enhances user engagement, ensuring correct pronunciation becomes an integral part of the learning process. In essence, the "English-Tamil Flashcards" app exceeds conventional language learning, emerging as a dynamic and user-centric educational tool that seamlessly blends technology, interactivity, and innovation.

2. LITERATURE SURVEY

In the field of language learning applications, this section explains the literature survey related to the proposed work is illustrated below.

Bimal Aklesh Kumar and Munil Shiva Goundar, in 2022, conducted an extensive exploration of the Mobile Language Learning (MLL) landscape. widespread adoption of app development, speech technology, and gamification in design, along with a prevalent reliance on usability testing for evaluation [1].

In 2019, Smith et al. conducted a comprehensive study examining the efficacy of different language-learning applications, with a particular emphasis on interactive flashcards [2]. Their research delved into the effectiveness of these tools, likely exploring aspects such as user engagement, retention, and overall language acquisition. The study contributes valuable insights to the field of language education, shedding light on the potential benefits of interactive flashcards in the

learning process. The findings may have implications for educators and learners seeking optimized language learning strategies based on technology.

In 2020, Brown et al. conducted a study scrutinizing the pedagogical influence of dynamic components, randomization techniques, and user-generated input in language learning, offering significant insights [3]. Their analysis likely explored how these features contribute to enhanced learning outcomes, potentially focusing on aspects such as engagement, adaptability, and personalized learning experiences. The research sheds light on the educational potential of incorporating dynamic elements and user-generated content into language learning applications, providing valuable guidance for educators and developers in optimizing language learning platforms.

In 2018, Chen et al, evaluated the influence of Graphical User Interface (GUI) design in language learning apps, emphasizing the role of user-friendly design in elevating engagement, and improving overall learning experiences and the context of audio pronunciation in language learning apps.[4].

In 2018, Patel et al, examined the integration of text-to-speech functionality in language learning technology, revealing how this feature enhances pronunciation and accessibility [5].

Kim et al. exploring 3D visualization's role in enhancing engagement and understanding, discusses the integration of three-dimensional visualization in educational tools, particularly in language learning applications, providing insights into its potential benefits in 2017[6].

Conducted by Ruo Wei Chen and Kan Kan Chan in 2019, this study delves into the effectiveness of Augmented Reality (AR) flashcards versus traditional paper flashcards in early childhood education. The research, involving 98 children aged 5-6, demonstrates that both methods significantly enhance vocabulary learning with no notable difference in effectiveness. Teachers observed children's enjoyment of AR activities but noted challenges in integrating AR flashcards into kindergarten settings [7].

These seminal works collectively contribute to the evolving landscape of language learning applications and provide a robust foundation for understanding their effectiveness and potential enhancements.

3. PROBLEM STATEMENT

Develop a Python-based English-Tamil Flashcard Application using Tkinter and external libraries. The application allows users to input a specified number of English word pairs, automatically translating them to Tamil. The flashcards are displayed on a Tkinter canvas, featuring buttons to play audio pronunciations in both English and Tamil. Additionally, users can explore a 3D visual representation of an image associated with the word. The program employs Google Translate API, gTTS for audio, and Pygame for 3D visualization. The application aims to enhance language learning through interactive flashcards and multi-sensory experiences.

4. DEFINITIONS

a. Google Translation API (googletrans): The Google Translation API, accessible through the googletrans Python module, enables developers to integrate Google's powerful translation capabilities into their applications. It allows text translation between various languages and supports both single sentences and larger pieces of text. The API is easy to use and provides a straightforward interface for language translation, making it a popular choice for multilingual applications and services.

b. Python Tkinter: Tkinter is the standard GUI (Graphical User Interface) toolkit that comes with Python. It provides a set of tools for creating desktop applications with graphical interfaces. Tkinter supports various widgets, allowing developers to design windows, buttons, menus, and more. Its simplicity and ease of use make it a preferred choice for developing basic desktop applications in Python.

c. Random Module: The random module in Python is a standard library module that provides functions for generating random numbers. Developers can use it to introduce randomness in their programs, simulations, or games. The module includes functions for generating random integers, floating-point numbers, and making random selections from sequences. Its versatility makes it useful in scenarios where unpredictability or variability is desired.

d. gTTS (Google Text-to-Speech): The gTTS module allows Python developers to easily convert text into speech using Google Text-to-Speech API. It's a simple and efficient tool for creating spoken audio from written content. Developers can generate speech files or directly play the output. This module is particularly useful for applications that require text-to-speech functionality, such as voice assistants or audio content generation.

e. Pygame: Pygame is a cross-platform set of Python modules designed for writing video games. It provides functionalities for handling graphics, sound, input devices, and more. Pygame simplifies the process of game development by abstracting low-level details and offering a high-level interface. It is widely used for both educational purposes and professional game development, making it a valuable tool for those looking to create 2D games in Python.

5. PROPOSED SYSTEM:

The proposed system begins with user initiation, where vocabulary pairs are inputted into the program. Subsequently, a Tkinter GUI is generated, featuring a flashcard display and control buttons for user interaction. Upon execution, the program autonomously selects and presents a flashcard, enriching the learning experience. Users have the option to engage further by listening to pronunciations and exploring a 3D image through dedicated buttons on the interface. The system ensures a continuous and interactive learning environment by incorporating a main event loop, facilitating ongoing user engagement until the decision to exit the application. This cohesive approach blends user input, graphical interface, dynamic content display, and interactive features to create an effective and engaging language-learning platform.

6. BILINGUAL FLASHCARD FRAMEWORK

Flashcards, featuring English words and their Tamil translations, expedite vocabulary acquisition and language proficiency. Tailored for test preparation, they enhance recall through repetition, ensuring readiness for assessments. Integrating cultural context, flashcards offer a holistic approach to learning Tamil. Portable and customizable, they enable on-the-go language study, catering to diverse learning styles. Active engagement with flashcards reinforces knowledge retention, deepening understanding of Tamil Language. As interactive aids, flashcards play a vital role in effective language learning strategies.

The phases of the proposed system are given below.

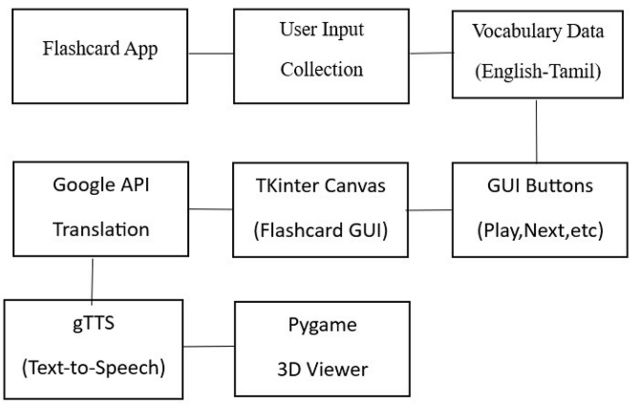


Fig.1: Language Flashcard App Structure

Phase 1: Initialization and User Input

The program starts by creating an instance of the Flashcard App class and initializing a Tkinter GUI. The user is prompted to enter the number of word pairs they want to input.

Phase 2: Vocabulary Building

In this interactive language-learning program, users input English words, initiating a dynamic process where each word undergoes translation to Tamil using the Google Translate API. The translated pair, comprising the original English word and its corresponding Tamil translation, is then systematically stored in a dictionary named 'vocab.' This approach ensures that users effortlessly build a bilingual vocabulary, seamlessly integrating the benefits of real-time translation into their learning experience.

Phase 3: GUI Creation

The program utilizes Tkinter to establish a canvas with a distinctive blue background, functioning as a visually appealing display for flashcards. Integrated within this canvas are strategically placed buttons, each designed to enhance the user experience. These buttons facilitate the pronunciation of both English and Tamil words, enabling users to listen and reinforce their auditory learning. Additionally, buttons for seamlessly transitioning to the next flashcard and unveiling a captivating 3D view further contribute to the interactive nature of the application. This thoughtful combination of a visually engaging canvas and strategically placed buttons creates an intuitive and comprehensive language-learning interface, ensuring a dynamic and user-friendly experience for learners.

Phase 4: Flashcard Display

The program randomly selects a word pair and displays it on the canvas as a flashcard.

Phase 5: Audio Pronunciation

Buttons trigger the generation and playing of audio pronunciations for the current word pair in English and Tamil.

Phase 6: 3D Viewer

A button triggers the display of a 3D view using Pygame, rotating an image loaded from a specified path.

Phase 7: Main Event Loop

The main event loop (root. main loop ()) manages the Tkinter GUI, handling user interactions and updating the display.

Phase 8: Program Termination

When the user closes the Pygame window or exits the Tkinter application, the program gracefully exits.

7. RESULT & DISCUSSION

The real-time application-based system helps to reduce the man's work. In addition to that, the yielded result is voice-based which helps the challenged person to seek knowledge towards a language.

7.1 User text

The user can give a word without a word constraint limit. This part translates an input to concern language output, which may help the challenged people to learn a language. From Fig. 2., the input can be given as much as the user needs and gets paired to yield a result. The particular language is given as an input and the result is obtained in a particular language, for example, the input is given in the English language and the output is generated in a Tamil language. The fig.3. shows the language to be entered as many users input given in the above input.

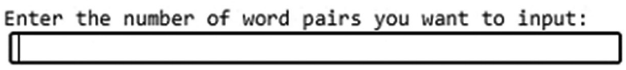


Fig.2: User input

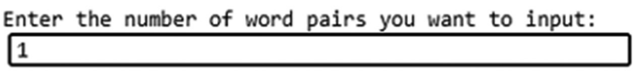


Fig.3: Enter the number of words

From Fig. 4., the English words (say) are given, as the user predefined the number of words to be paired.

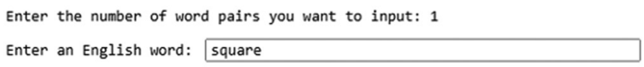


Fig.4: Enter English words to translate

7.2 Flashcard output

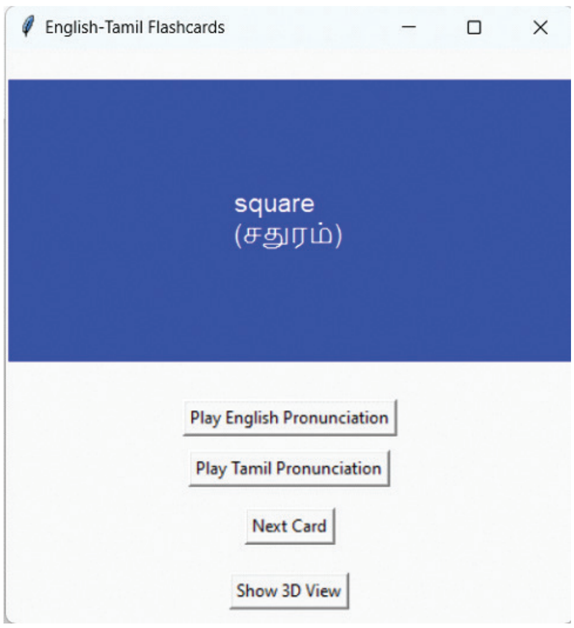


Fig.5: Translated Flashcard

The above Figure 5 explains the Flashcard, comprising the provided word and its translation. Additionally, this application is integrated with audio pronunciation for both English and Tamil languages, aiding in a more precise understanding of words. The transition to the next card can be initiated, and further, a 3D image of the given word is provided.

7.3 Audio Pronunciation

In Figures 6 and 7, the audio pronunciation feature is elucidated, showcasing the application's utilization of the gTTS (Google Text-to-Speech) module. This module serves a pivotal role in enhancing comprehension and refining accents through the generation of audio for both English and Tamil words. The 'play_english_audio' and 'play_tamil_audio' functions encapsulate the process, leveraging gTTS to convert text into clear and articulate speech. The generated audio files are intelligently saved and subsequently played back, providing users with a dynamic auditory dimension to their language-learning experience.

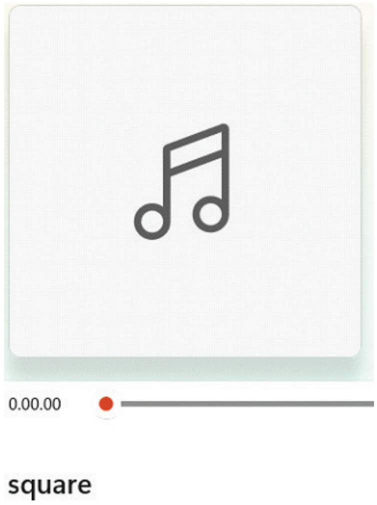


Fig.6: English Audio Pronunciation

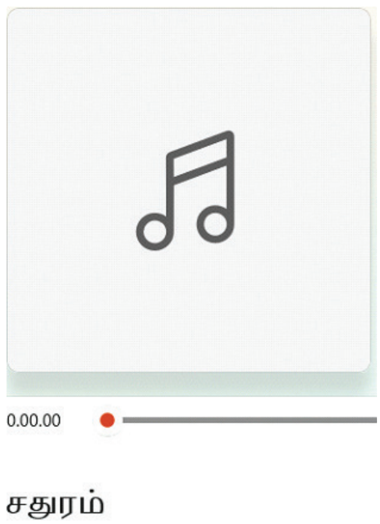


Fig.7: Tamil Audio Pronunciation

7.4 3D Visualization

Figure 8 captures the output of a 3D image associated with the provided word, offering users a visually immersive experience that enhances their conceptual understanding. This feature goes beyond traditional language learning methods, providing a dynamic and engaging way for users to interact with and comprehend words in a spatial context. The visual representation stimulates excitement and curiosity, fostering a positive learning environment by combining visual and linguistic elements. This innovative approach not only aids in language acquisition but also promotes a deeper and more memorable understanding of the presented vocabulary.

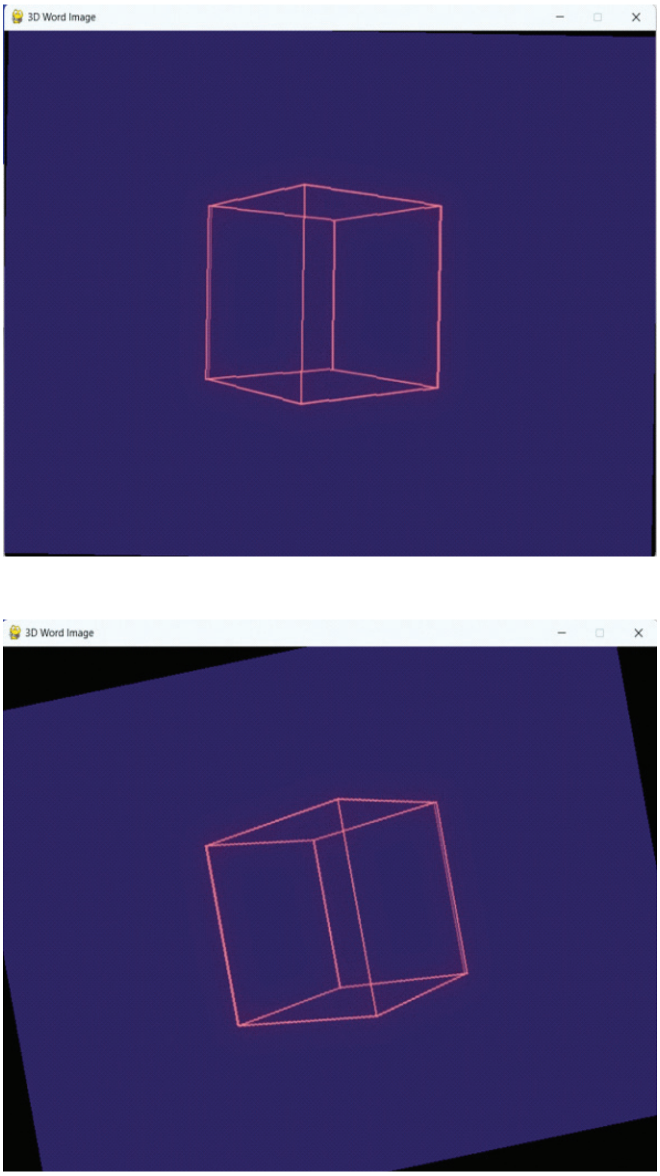


Fig.8: The rotated view of the 3D image

The real-time application-based system presented here offers a significant reduction in manual effort, particularly benefiting individuals facing language learning challenges. The voice-based output further caters to those with specific learning needs. The user-friendly system allows unlimited input, facilitating language learning for challenged individuals. The Flashcard component depicts this by pairing user-provided words with translations, presenting an inclusive approach to language acquisition. Additionally, the integration of audio pronunciation using gTTS underscores the application's commitment to enhancing comprehension and refining accents. The dynamic auditory dimension, coupled with a 3D viewer feature, collectively provides a comprehensive language learning experience. Looking forward, potential enhancements include language support expansion, gamification elements, and adaptive learning pathways through machine learning. Iterative improvements driven by user feedback will guarantee the continued efficacy and pertinence of the application within the dynamic realm of language education.

8. CONCLUSIONS

The English-Tamil Flashcard application stands at the forefront of educational innovation, providing an integrated language learning experience. Its dynamic generation of flashcards, coupled with user input functionality, encourages personalized journeys for learners. The commitment to auditory learning, evident through audio pronunciation support, ensures a comprehensive language immersion. The inclusion of a 3D viewer feature adds a visually stimulating dimension to vocabulary exploration. Looking ahead, future developments could include expanding language support, incorporating gamification elements for enhanced engagement, and integrating machine learning algorithms for adaptive learning pathways. Furthermore, continuous updates and user feedback mechanisms will be essential to refine and optimize the application's features, ensuring it remains a cutting-edge and effective tool in the ever-evolving landscape of language education.

REFERENCES

[1] Bimal Aklesh Kumar, Munil Shive Goundar, "Developing Mobile Language Learning Application: A systematic" Literature Review in 2022.

[2] Smith, "Effectiveness of Various Language Learning Applications: A Comprehensive Study" in 2019.

[3] Brown, "Analyzing the Pedagogical Impact of Interactive Flashcards in Language Learning" in 2020.

[4] Chen, "Evaluating GUI's Influence on Engagement and Learning Outcomes in Language Learning Apps" in 2018.

[5] Patel, "Examining the Integration of Text-to-Speech Functionality in Language Learning Technology" in 2018.

[6] Kim, "Exploring 3D Visualization's Role in Enhancing Engagement and Understanding in Educational Tools" in 2017.

[7] Ruo Wei Chen, Kan Kan Chan, "Using Augmented Reality Flashcards to learn vocabulary in early childhood Education" in 2019.

[8] Garcia, "Investigating the Significance of Audio Pronunciation in Language Learning Apps" in 2019.

[9] Wang, "Exploring the Benefits of User Cu Tomization in Educational Apps" in 2016.

Singapore's Tamil Digital Technologies: A Diaspora Pathseeker

Arun Mahizhnan & Nara Andiappan

ABSTRACT

Singapore has been an early adopter of new digital technologies because of economic imperatives and governance requirements. The government has also harnessed these technologies for educational and cultural purposes, among other needs. These needs, in combination with the government's multilingual policies, enshrined in the Constitution, have given the Tamil language in Singapore an extraordinary opportunity to engage Tamil-related digital technologies. This paper will address some key elements of Tamil Digital Technology (TDT) in the service of education, media and governance in Singapore.

The challenges for developing TDTs in Singapore are formidable. Apart from financial constraints, which are normal, there are severe shortages of talent in TDT. Singapore, therefore, needs collaboration with the outside world. Tamil Nadu, given its bounty of talents and resources, can lead the way for the whole diaspora. But it can only be sustained if the effort is based on professionalism, efficiency and equity among the partners.

Arun Mahizhnan,
Special Research Adviser at the Institute of Policy Studies(IPS) at the Lee Kuan Yew School of Public Policy in the National University of Singapore.
Email: arunmahizhnan@outlook.com

Nara Andiappan,
Founding member of Centre for Singapore Tamil (CSTC).
Email: naraandiappan@gmail.com

1. INTRODUCTION

The Singapore government has been an early adopter of new digital technologies because of economic imperatives and governance requirements. The government has also harnessed these technologies for educational and cultural purposes, among other needs. Combined with these needs, the government's multilingual policies, enshrined in the Constitution, have given Tamil an extensive platform for engaging Tamil-related digital technologies. This paper will address some key elements of Tamil Digital Technology (TDT) in the service of three specific domains: education, media and governance.

In drawing up the first Information Technology MasterPlan – IT2000 – in 1996, the government laid the foundation for a National Information Infrastructure (NII). It declared that "Singapore will be among the first countries in the world to have a national information infrastructure"[1] and that it was the foundation for transforming Singapore into an "Intelligent Island." The plan aimed to build a pervasive network interconnecting computers in virtually "every home, office, school and factory."

The IT2000 was succeeded by two other 10-year masterplans, the Intelligent Nation 2015 (iN2015) launched in 2005 and the Digital Connectivity Blueprint (DCB) launched recently in June 2023[2]. While the IT2000 masterplan focused on the availability of internet technologies, iN2015 sought to improve the accessibility of these technologies through a variety of digital services and businesses to reach the masses. Building on these foundational plans, DCB is designed to empower the community to develop Singapore as a 'smart nation'.

While focusing on general technological capabilities, the government has kept in mind that the different segments of society should not be unduly or unfairly disadvantaged. Inclusivity and equity have been the underlying thrusts of its digital efforts.

1. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/archived/ida/press-releases/1996/20050726143657
2. https://www.imda.gov.sg – Infocomm Media Development Authority of Singapore

Language communities have thus been given special attention, underpinned by the multilingual policies of the government.

For those not familiar with Singapore's ethnic and linguistic landscape, Indians make up roughly nine percent of the population while Chinese and Malays constitute, respectively, about 75 percent and 15 percent. Whereas people are identified based on their ethnicity, when it comes to the designation of official languages, only three are recognised for the three major ethnic groups – Mandarin for the Chinese, Malay for the Malays and Tamil for the Indians. Other languages of all these ethnic groups are indeed used by them in Singapore but do not enjoy official status.

Thus, Tamil has been given an official position in various language initiatives of the government and when it came to digital technologies, Tamil was included from the very beginning of national efforts. Currently, the most eagerly anticipated outcomes are related to Tamil AI, a topic we shall address later.

2. EARLY TAMIL DIGITAL TECHNOLOGIES

When information and communication technologies were spreading around the world, and Singapore became an early adopter, Tamil was duly ensconced on the multilingual web platform the government was building. Tamil appeared on the Singapore web for the first time on 27 October 1995[3]. The key individuals who made this possible were Dr Tan Tin Wee, an ethnic Chinese technologist and Mr N Govindasamy, a Tamil educator.

While Tamil had an auspicious start on our web, it was not something the tradition-bound Tamil leadership took to readily or easily. Though I (the first author Arun Mahizhnan) was not part of the Tamil language leadership in Singapore, I was brought in by Dr Tan and some political leaders to steer the Tamil internet drive in Singapore. I was extremely fortunate to have the strong support of Dr Tan and Muthu Nedumaran, the only full-time techie in Singapore at that time, who were deeply involved with Tamil internet globally. Though Malaysian by birth and citizenship, he benefited greatly by working in Singapore and he paid back some of the debt by immersing himself in Singapore Tamil Internet. If not for him and Dr Tan, the first ever worldwide Tamil internet organisation called the International Forum for Information Technology in Tamil, INFITT for short, would not have been founded in Singapore in 2000. Prof M Anandakrishnan, the then IT Advisor to the Tamil Nadu Government, was elected the founding Chairman of INFITT. He was instrumental in persuading most members to set up the INFITT Secretariat in Singapore and to appoint me as its first Executive Director. While these individuals were the visible leaders of INFITT, there was the invisible hand of the Infocomm Development Authority, which provided all the governmental and financial support for the world's biggest Tamil internet conference till then, and the establishment of the INFITT Secretariat in Singapore. And my fellow author, Nara Andiappan, was the first paid employee of INFITT – hired by IDA and loaned to INFITT.

The reason why we mention this particular development in Singapore is to highlight three key aspects of international collaboration: First, the environment in which an international organisation is based has to be conducive to international collaboration. Singapore was already a global city, with global connections and exposure to a global culture. Second, there must be an attitude of humility and a hunger for learning. Perhaps because our own neighbour had knocked on our head and told us that we are nothing but a little red dot on the map, we have always been acutely aware of our size – rather, the utter lack of it. And we have always been willing to learn from others as we don't have massive resources or even a huge population. Our survival depends on our learning from others. This is why, even today, the final message we will leave you with relates to that hunger for learning. Third, perhaps the most important aspect is the abiding government support – even when it is only a minority community that is affected, the Singapore government takes a serious and sustained interest in supporting Tamil-related technologies. In terms of economic and political impact, Tamil is not a critical factor but as a multicultural nation, Singapore sees the benefits of language-related technologies. Their impact on governance, education, and media is huge and lasting.

3. TAMIL DIGITAL EDUCATION

Let us now turn to the three domains which form the focus of this paper. Perhaps the single most important domain where TDTs can make the greatest impact is Tamil education. The Singapore Tamil education field involves thousands of Tamils – as students and teachers. Currently, about 25,000 students are studying Tamil and 800 teachers teaching Tamil in Singapore.

Late last year, the Ministry of Education (MOE) unveiled the "EdTech Masterplan 2030" to harness the transformative potential of technology in education. It outlines how schools can better leverage technology to enhance teaching and learning. This masterplan is only the latest edition of many educational masterplans of the Singapore government over the years. The following is the gist of the 2030 masterplan that is relevant to Tamil education:

3. https://www.infitt.org/ti1999/papers/naago1998.pdf

The EdTech Masterplan 2030 aims to transform education through technology, focusing on four outcomes. They are namely, students who are digitally empowered, teachers who are technologically adept and collaborative learning designers, schools which are an intelligent, responsive, digitally equipped learning environment, and a system that is a networked EdTech ecosystem.

Under the outlined master plan, opportunities will be presented for Tamil language teachers to enhance their expertise in utilising technology, including the effective integration of Artificial Intelligence.

Student Learning Space

The Student Learning Space (SLS) portal, introduced by MOE in 2018, is aimed at revolutionising the learning experiences of Singaporean students through the strategic use of technology. This student-centric portal caters comprehensively to all subjects, including Tamil Language. Numerous Singaporean Tamil teachers actively utilise the SLS to design ICT-integrated lessons, sharing their expertise by publishing lessons for their fellow Tamil teachers. There is significant interest among Tamil teachers to enhance their proficiency in utilising open tools for language teaching, such as Thinklink, Kahoot, Mentimeter, Quizlet, Decktoys, Edpuzzle, Nearpod, Padlet, and others. The adoption of digital spaces for learning enables students to engage in diverse learning modes, fostering both self-directed and collaborative learning.

SkillsFuture for Educators: Effective Use of e-Pedagogy

The Learn for Life movement, initiated by MOE in 2020, encompasses the SkillsFuture for Educators (SFEd) roadmap, underscoring the importance of continuous professional development for teachers. Accordingly, Tamil teachers, like their counterparts in other subjects, employ digital technologies to enhance and personalise students' learning experiences. They actively design e-Pedagogy integrated lessons on the SLS platform and various web platforms, contributing to the effective teaching of Tamil,

However, the use of Tamil beyond the classroom, especially at home, has been on the decline. This is a fate shared by other mother tongues such as Mandarin and Malay. English is on the ascendance in many homes. The government and the community, together, have been exploring ways and means to encourage the use of Tamil outside the classroom. Digital games in Tamil are proving to be an engaging tool but hardly enough to reverse the trend in any significant way. Many more innovative and engaging TDT-based tools are needed.

In addition, Singapore has been suffering from a shortage of Tamil teachers for some time. MOE's strategies to overcome this problem have faced considerable challenges. Perhaps the time has come to change tacks and look at non-conventional strategies. Online teaching as a supplement to face-to-face teaching is worth considering. In fact, COVID forced our entire teaching system to adopt online teaching for a period. Many teachers, originally reluctant to accept that online teaching can be effective, found it to be so. While it will not be a complete replacement for face-to-face teaching, it can certainly be an additional tool. Top universities in the world have been delivering online courses for years now. Online Tamil teaching at the secondary school and junior college level may be worth exploring.

There is another reason why we emphasise this particular point. Throughout the Tamil diaspora, there is a critical and unbreachable shortage of capable Tamil teachers on-site. For those Tamil communities, online teaching may not only be a supplementary tool but the primary tool. The Tamil Virtual Academy, which has convened this great conference, has already been playing a role in addressing this problem and we hope it will join hands with other Tamil Nadu institutions to offer much more. Tamil Nadu has the talent and resources to lead this initiative.

4. TAMIL DIGITAL MEDIA

Singapore's Tamil media history dates back to 1875, when Singapore's first Tamil newspaper, Singai Varthamani, was first published. As for broadcasting, Tamil radio broadcasts began in the 1930s. Television came to Singapore in 1963, with Tamil TV as an integral part. From the Gutenberg Press to the Internet, many communication technologies have had their impact on world media, including Singapore's. However, the introduction of TDTs in Tamil media has opened the doors to an entirely new world of information, education and entertainment.

Two media companies provide most of the Tamil mass media content in Singapore. The SPH Media Trust has a history starting in 1845, barely 25 years after the British founded Singapore as a trading post to be managed from Calcutta. It has a long and illustrious history. SPH Media publishes newspapers in all four official languages, including the only surviving Tamil daily called Tamil Murasu. Tamil Murasu was started independently by the preeminent Tamil community leader Tamilavel Ko Sarangapany in 1935 but eventually ended up in the hands of the largest print publisher. But what we could not do and could not even imagine doing during Sarangapany's time -- that is reading Tamil Murasu daily from any corner of the earth -- you can now do, thanks to TDTs. We are sure Tamilavel would be pleased. Murasu went online as an e-paper in 2017 and is updated as many times a day as necessary. It

is still published as a hard copy paper as well. SPH Media, which launched a free mobile app for the daily in October 2023, is trying to incorporate whatever new technologies are available, and the Tamil division shares the benefits of these technologies wherever and whenever it can adapt them to the Tamil language.

The other major media organisation in Singapore is Mediacorp, the biggest broadcaster and local content provider in all four official languages. Broadcasting had mostly been a government operation for decades until it was corporatised in 1980. Today, we have Tamil radio, television and online content 24 hours a day through various channels and delivery systems, thanks to digital technologies. Tamil news went online in 2015, providing consumers with rapid updates every day. These developments were made possible not merely by conducive market conditions or profit motives but also by government policy. The Tamil media output by Mediacorp and SPH Media is subsidised by government funds. This is why, we the Tamils of Singapore have to treat our mass media with a great deal of responsibility. We cannot afford to squander these precious assets of the community in a cavalier manner.

5. TAMIL DIGITAL GOVERNANCE

As a whole, the government, through its key agencies, has been doing the heavy lifting in enabling TDTs in Singapore. Often it is the pioneer in such developments. We would like to focus on four specific areas in which the Singapore government has deployed TDTs for the benefit of the public.

National Library Board

Perhaps, among all the government services provided in Tamil, the most widely known within the community is the National Library Board (NLB) – with the obvious exception of the MOE services, which we mentioned earlier. NLB is the central knowledge-building institution in all four languages under one roof. It has been serving the Tamil community since colonial days, through its predecessor organisations. It was among the earliest government agencies to adopt Tamil technologies. Tamil catalogues were digitised in 2002/2003, probably among the first in the Tamil diaspora. Its digital collection of materials, which started in 2006, has reached more than 7,000 items by 2023. One of the most noteworthy accomplishments was the digitisation of 50 years of Singapore Tamil literary publications in 2015, to commemorate the 50th anniversary of Singapore’s Independence. It was a pioneering effort in any language in Singapore. Interestingly, it was a project which a group of Tamils proposed and carried out in collaboration with NLB. The same community group is now working with NLB on creating the first-ever digital encyclopedia of Singapore Tamils. Such Tamil digital resources would not only be available to Singaporeans but to anyone, anywhere, anytime and for free.

Digital government services

Since 2018, the government has begun offering commonly used government services in multiple languages including Tamil. With this facility, a citizen can use her identity app (SingPass) in Tamil to access numerous government services and websites. Anyone can simply book a medical appointment at a polyclinic, find information on government programmes, learn new digital skills or apply for housing services – all in Tamil.

Public communications

The Ministry of Communications and Information (MCI) launched the SG Translate Together portal in 2022. The idea was to harness the collective wisdom of the citizenry in enhancing translated scripts in Chinese, Malay and Tamil for the government's use in its daily operations. SG Translate, a neural machine translation engine, can use localised translation data and generate translation of texts between English and other official languages. This allows government agencies to produce more reliable content that is attuned to the local context and culture.

Emerging technologies

Currently, the most widely anticipated outcomes of Tamil Digital Technologies are related to Tamil AI. Singapore recently launched an ambitious US$52 million AI initiative to develop Southeast Asia’s first large language model (LLM) ecosystem. However, this initiative’s success is largely dependent on the availability of billions of data points amassed from the digital content of each regional language. Tamil is automatically included due to its official position in Singapore. Additionally, the Infocomm Media Development Authority is also developing new benchmarks and tools for an AI governance framework. This initiative aims to identify gaps in policies and plans related to AI where the AI tools are challenged for their assumptions in Singapore's multi-lingual context, including that of Tamil. Tamil AI is, understandably, the least developed of the digital technologies today, but the trajectory and the eventual benefits would be bigger and longer lasting than most past technologies. However, one thing that is clear to us is that more than anything else in the past, we need more collaboration and cooperation within the Tamil world.

6. TDT COLLABORATION

The one word, the one idea we want to leave you with is COLLABORATION. Singapore seeking

collaboration is inevitable – as we noted earlier. However, it is our understanding that the new digital Tamil technologies would move rapidly, smoothly and for the benefit of all if we all collaborate. As the Tamil Internet story amply demonstrates, we achieved much when we worked together. And we started to decline when our collaboration weakened.

Tami Nadu is the North Star of the Tamil diaspora. In terms of knowledge, talent and resources, Tamil Nadu is unmatched by any other country in the world. That preeminence sometimes could lead to a certain aloofness; it could encourage the go-it-alone tendency. But, if Tamil Nadu were to engage and involve the diaspora Tamils in developing and deploying TDTs, all of us stand to gain much. In this regard, we note with great pleasure that the TN Government has already established mechanisms to harvest the diaspora talent and technologies.

For our part – our tiny little part – Singapore is willing and able, and indeed, needs to collaborate with anyone, especially Tamil Nadu. We can start small and then ramp up fast. "Akalakkaal" (biting off more than we can chew) doesn't suit us.

While everyone would agree collaboration is a good thing, experience teaches us that it can only be successfully sustained if the effort is based on professionalism, efficiency, and equity among the partners. We hope all our partners will bear these qualities in mind as we embark on a journey that will surely take us to new horizons, and new dawns.

7. CONCLUSION

New Tamil Digital Technologies are emerging in good numbers but still have a long way to go to match what is available in English and some other languages. It is also the case that with greater integration of the Tamil diaspora, developments in one corner of the world could benefit other parts. That is why a conference of this nature is profoundly important for the whole diaspora. We hope this initiative will continue into the future to help all of us reach much greater heights.

ROLE
OF
TAMIL
IN
SPATIAL
COMPUTING

Integrating Spatial Computing for Promoting Tamil Language

Srisivasubbramanyan BS, Gayathri P, Dr S Kanaga Suba Raja

ABSTRACT:

This research seamlessly integrates Tamil language, scriptures, and literature into spatial computing. Preserving and promoting Tamil cultural heritage. Leveraging advanced algorithms, it bridges the gap between traditional texts and contemporary experiences Key strategies include Natural Language Processing (NLP) for Tamil, employing tokenization, parsing, and Named Entity Recognition (NER) to deconstruct and analyse texts. Tailored Text-to-Speech (TTS) systems enrich the auditory experience, while Computer Vision using Optical Character Recognition (OCR) digitizes and analyses Tamil manuscripts. Spatial Audio and Gesture Recognition enhance user engagement, and Augmented Reality (AR) overlays digital representations onto real-world objects. Spatial Data Visualization illustrates connections in Tamil literature, and Machine Learning algorithms provide personalized recommendations. Collaborative Filtering fosters community engagement, addressing cultural sensitivity and accessibility challenges through user interface design. Digital Preservation Techniques, like metadata tagging, enhance searchability. This interdisciplinary effort aims to underscore Tamil's transformative role in spatial computing, fostering global appreciation for linguistic diversity and cultural heritage

Srisivasubbramanyan BS
Department of CSE, SRM Institute of Science and Technology, Trichy, India. Email: sivasubbu421@gmail.com
Gayathri P
Department of CSE, SRM Institute of Science and Technology, Trichy, India. Email: gayukeerthy@gmail.com
Dr S Kanaga Suba Raja
HOD Department of CSE, SRM Institute of Science and Technology, Trichy, India. Email: skanagasubaraja@gmail.com

I. INTRODUCTION:

In an era where technology and cultural preservation converge, this research delves into the revolutionary integration of the Tamil language, its scriptures, and literature into the immersive realm of spatial computing. The overarching goal is to preserve and promote the intricate tapestry of Tamil cultural heritage, fostering a connection between traditional narratives and contemporary experiences through the application of cutting-edge algorithms. At the heart of this study are advanced strategies that leverage Natural Language Processing (NLP) techniques tailored specifically for Tamil. These encompass processes such as tokenization, parsing, and Named Entity Recognition (NER), enabling a nuanced deconstruction and analysis of Tamil texts. The auditory facet of the Tamil literary experience undergoes a transformation with the implementation of Tailored Text-to-Speech (TTS) systems, providing a dynamic vocalization of age-old narratives that adds a novel dimension to the user experience. The research further embraces Computer Vision techniques, particularly Optical Character Recognition (OCR), to digitize and analyse Tamil manuscripts. This not only contributes significantly to the preservation of cultural artifacts but also facilitates the seamless transition of tangible, traditional texts into the digital sphere.

To heighten user engagement, the study incorporates Spatial Audio and Gesture Recognition. Spatial Audio techniques create an immersive auditory environment for users to experience 3D renditions of Tamil literary works, enhancing the overall engagement. Gesture Recognition algorithms empower users to interact intuitively with spatial representations of Tamil literature, introducing a novel and accessible form of engagement. Augmented Reality (AR) plays a pivotal role by overlaying digital representations onto real-world objects, creating a tangible and interactive bridge between the virtual and physical dimensions of Tamil literature.

Spatial Data Visualization serves as a graphical conduit, illustrating the intricate connections within Tamil literature. This visualization not only enhances the understanding of literary relationships but also serves as a dynamic tool for users to explore and appreciate the depth of Tamil cultural heritage. Machine

Learning algorithms add a personalized touch to the user experience by offering tailored recommendations based on individual preferences. Collaborative Filtering ensures community involvement in the preservation of Tamil cultural heritage, fostering a collective and participatory approach. Addressing nuanced challenges related to cultural sensitivity and accessibility, the research emphasizes thoughtful user interface design. This ensures that the spatial computing applications are not only technologically advanced but also culturally sensitive and inclusive.

Digital Preservation Techniques, including metadata tagging, are implemented to enhance the searchability and categorization of Tamil literary content. This not only aids in the organization and retrieval of information but also contributes to the long-term preservation of Tamil cultural artifacts in the digital landscape.

In essence, this interdisciplinary research endeavour aims to illuminate the transformative role that Tamil can play in shaping the future of spatial computing. By seamlessly integrating cultural heritage with cutting-edge technology, the study seeks to foster a global understanding and appreciation for linguistic diversity, ensuring that Tamil's rich tapestry is not only preserved but also becomes an integral part of the evolving digital landscape. The subsequent sections of this paper will delve into the specific methodologies, challenges, and outcomes of this pioneering effort, providing a comprehensive exploration of the transformative potential inherent in the fusion of Tamil culture and spatial computing.

1.1 Motivation of the Proposed Methodology

- To Safeguard Tamil heritage by digitizing traditional texts and linguistic nuances in spatial computing.
- Seamlessly integrate age-old narratives into interactive formats, bridging traditional Tamil texts with spatial computing.
- Employ spatial audio, gesture recognition, and AR to make Tamil literature interactive and accessible.

1.2 Contribution of the Proposed Methodology

- Digitalization and preservation of Tamil heritage, enriching accessibility and cultural understanding in spatial computing.
- Through the Proposed model we have introduced spatial audio, AR, and gestures, providing an immersive and interactive exploration of Tamil literature.
- Personalized recommendations and user-friendly interfaces, ensuring diverse audiences engage effectively.
- Utilization of advanced algorithms for accurate cultural representation, addressing challenges and ensuring authenticity.

The proposed method is Developed in the Google Collab and the results are compared with the existing approach. The comparison results demonstrate that the suggested method outperforms the current method.

1.3 Structure of the Proposed Methodology

Chapter 1 deals with the introduction part. Chapter 2 dwells into related work. The proposed method is given in Chapter 3. In Chapter 4 (Results and discussions) The implementation's specifics and outcomes are shown. Chapter 5 is about the Conclusion and Future Work.

II. RELATED WORKS:

In the realm of cultural heritage preservation and immersive experiences, our project builds upon a foundation laid by several notable related works. One pivotal area of exploration has been the application of spatial computing techniques to cultural heritage. Research by Patel et al. [1] introduced the integration of spatial computing for preserving and promoting Tamil cultural heritage, providing a conceptual backdrop for our project. Their work underscores the importance of leveraging advanced algorithms and immersive technologies within the cultural preservation domain. Advancements in natural language processing (NLP) for cultural heritage analysis has also paved the way for our research. Kim and Gupta [2] delved into the intricacies of applying NLP to analyse Tamil literature, establishing a connection between linguistic analysis and cultural preservation. While this work offers valuable insights into textual understanding, our project takes a step further by integrating spatial computing elements, enhancing user experiences beyond linguistic comprehension. The exploration of augmented reality (AR) applications in cultural heritage, as surveyed by Chen and Kumar [3], provides inspiration for our marker less AR integration. However, our project distinguishes itself by tailoring AR experiences specifically for Tamil literature, offering a more comprehensive and culturally sensitive approach.

Machine learning in the context of cultural heritage recommendation systems has been explored by Li and Park [4]. While their work focuses on content recommendations, our collaborative filtering techniques foster community engagement, ensuring a collective endeavour reflective of the Tamil community's insights and aspirations. Lastly, the application of digital preservation techniques, coupled with metadata tagging, as showcased in the study by Lee et al. [5], contributes to our emphasis on searchability and categorization within the spatial application. Our project extends this concept

by integrating Dijkstra's Algorithm for optimized pathfinding, further enriching the user experience within the spatial environment. The related works in spatial computing, NLP, AR, machine learning, and digital preservation provide essential insights. Our project synthesizes these elements, offering a transformative spatial application that not only preserves Tamil cultural heritage but also fosters a global appreciation for the linguistic diversity and historical richness embedded in Tamil literature. Through this interdisciplinary effort, we aim to illuminate the transformative role that Tamil can play in shaping the future of spatial computing, making it an integral part of the global digital landscape.

III. PROPOSED MODEL:

The transformative journey within the Google Collab environment unfolds seamlessly, as the process aims to intricately weave the rich tapestry of Tamil language, scriptures, and literature into the immersive realm of spatial computing. It begins with the input of Tamil text, which serves as the foundation for subsequent processes. The text undergoes a linguistic analysis involving tokenization and Named Entity Recognition (NER). Tokenization dissects the complex text into individual words or tokens, while NER identifies and categorizes entities such as names, locations, and cultural references. This linguistic dissection lays the groundwork for a deeper understanding of the cultural nuances embedded in the Tamil language. The journey then progresses to the synthesis of an auditory dimension through a custom Text-to-Speech (TTS) system designed for Tamil. This TTS system dynamically vocalizes age-old narratives, epics, and poetic expressions, bringing a melodic and authentic resonance to the spatial application. The synthesized audio output is played, providing users with an immersive auditory experience and transcending the limitations of written text. Taking a significant turn into the realm of spatial computing, a Spatial Application is introduced. Dijkstra's Algorithm, a pathfinding optimization tool, is integrated into the Spatial Application to optimize paths or connections within the spatial representation. This algorithm plays a pivotal role in enhancing navigation or interaction within the application, contributing to a more seamless and efficient user experience. Spatial Audio techniques and Gesture Recognition algorithms are then implemented within the Spatial Application. Spatial Audio provides a three-dimensional representation of Tamil literary works, enriching the auditory experience. Simultaneously, Gesture Recognition empowers users to interact intuitively with spatial representations of Tamil literature, introducing a novel and accessible form of engagement. This combination bridges the gap between the virtual and physical dimensions, enhancing user immersion. The Integration of technology and reality is brought to life through Marker less Augmented Reality (AR) within the Google Collab environment. These functionality overlays digital representations of Tamil texts onto real-world objects, creating a tangible and interactive bridge between the virtual and physical dimensions. This immersive experience captivates users, ensuring a seamless blend of traditional and contemporary elements within the application. Spatial data visualization techniques further enhance the user experience by illustrating intricate connections within Tamil literature. Graphical representations serve as a navigational guide, providing users with insights into the relationships between different literary works, authors, and historical periods in Tamil literature. This visual aspect enriches the user experience, facilitating a deeper understanding of the cultural tapestry.

The application prioritizes user interface design considerations, addressing nuanced challenges of cultural sensitivity and accessibility. By ensuring inclusivity and cultural relevance in the interface, the goal is to create an application that resonates with users across varying levels of familiarity with Tamil literature. The comprehensive flow of this transformative methodology in the Google Collab environment is a symphony of advanced algorithms and spatial computing techniques. It not only preserves Tamil cultural heritage but also transcends boundaries, fostering global appreciation for the linguistic diversity and historical richness embedded in the tapestry of Tamil literature. Through this interdisciplinary effort, the application aims to illuminate the transformative role that Tamil can play in shaping the future of spatial computing, making it an integral part of the global digital landscape.

3.1 Input Processing: Foundation for Transformation

The transformative journey commences with the foundational step of inputting Tamil text into the Google Collab environment. This textual content, whether drawn from scriptures, literature, or cultural artifacts, serves as the raw material for subsequent processes. This process is achieved by OCR

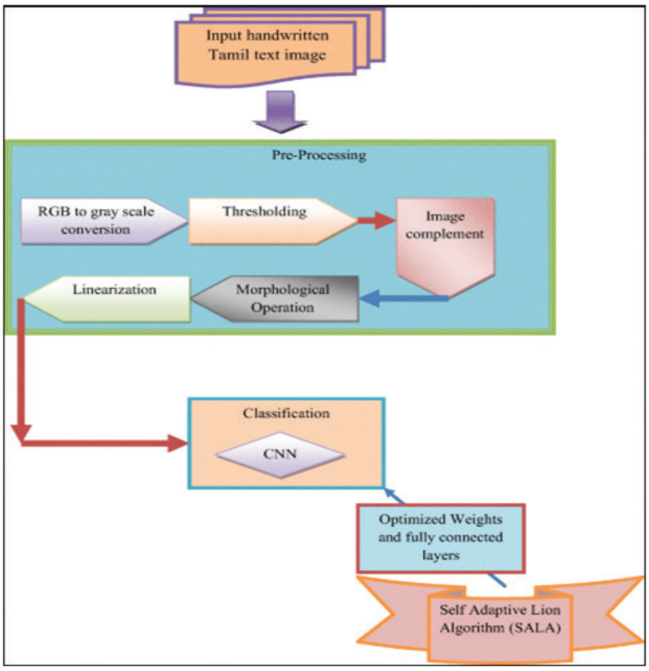


Image 1: Representing the process of OCR Optical Character Recognition

3.2 Linguistic Analysis: Tokenization and Named Entity Recognition (NER)

The text undergoes a comprehensive linguistic analysis, beginning with tokenization, where the complex textual content is dissected into individual words or tokens. Following this, Named Entity Recognition (NER) is applied to identify and categorize entities such as names, locations, and cultural references. This linguistic dissection lays the groundwork for a nuanced understanding of the cultural nuances embedded in the Tamil language.

3.3 Auditory Synthesis: Custom Text-to-Speech (TTS) System

After linguistic analysis, the journey progresses to the synthesis of an auditory dimension through a custom Text-to-Speech (TTS) system designed explicitly for Tamil. This system dynamically vocalizes age-old narratives, epics, and poetic expressions, imparting a melodic and authentic resonance to the spatial application. The synthesized audio output transcends the limitations of written text, offering users an immersive auditory experience.

3.4 Spatial Computing Introduction: Integration of Dijkstra's Algorithm

Taking a significant turn into the realm of spatial computing, a Spatial Application is introduced. Dijkstra's Algorithm, a pathfinding optimization tool, is seamlessly integrated to optimize paths or connections within the spatial representation. This algorithm plays a pivotal role in enhancing navigation or interaction within the application, contributing to a more seamless and efficient user experience.

3.5 Audio and Gesture Interaction: Enhancing User Immersion

Spatial Audio techniques and Gesture Recognition algorithms are implemented within the Spatial Application. Spatial Audio provides a three-dimensional representation of Tamil literary works, enriching the auditory experience. Simultaneously, Gesture Recognition empowers users to interact intuitively with spatial representations of Tamil literature, introducing a novel and accessible form of engagement. This combined approach bridges the gap between the virtual and physical dimensions, enhancing user immersion.

3.6 Marker less Augmented Reality (AR): Blending Technology with Reality

The integration of technology and reality is brought to life through Marker less Augmented Reality (AR). These functionality overlays digital representations of Tamil texts onto real-world objects, creating a tangible and interactive bridge between the virtual and physical dimensions. This immersive experience captivates users, ensuring a seamless blend of traditional and contemporary elements within the application.

3.7 Spatial Data Visualization: Illuminating Connections

Spatial data visualization techniques further enhance the user experience by illustrating intricate connections within Tamil literature. Graphical representations serve as a navigational guide, providing users with insights into the relationships between different literary works, authors, and historical periods in Tamil literature. This visual aspect enriches the user experience, facilitating a deeper understanding of the cultural tapestry.

3.8 User Interface Design: Culturally Relevant and Inclusive

The application prioritizes user interface design considerations, addressing nuanced challenges of cultural sensitivity and accessibility. By ensuring inclusivity and cultural relevance in the interface, the goal is to create an application that resonates with users across varying levels of familiarity with Tamil literature.

3.9 Transformation

The comprehensive flow of this transformative methodology within the Google Collab environment represents a symphony of advanced algorithms and spatial computing techniques. It not only preserves Tamil cultural heritage but also transcends boundaries, fostering global appreciation for the linguistic diversity and historical richness embedded in the tapestry of Tamil literature. Through this interdisciplinary effort, the application aims to illuminate the transformative

role that Tamil can play in shaping the future of spatial computing, making it an integral part of the global digital landscape.

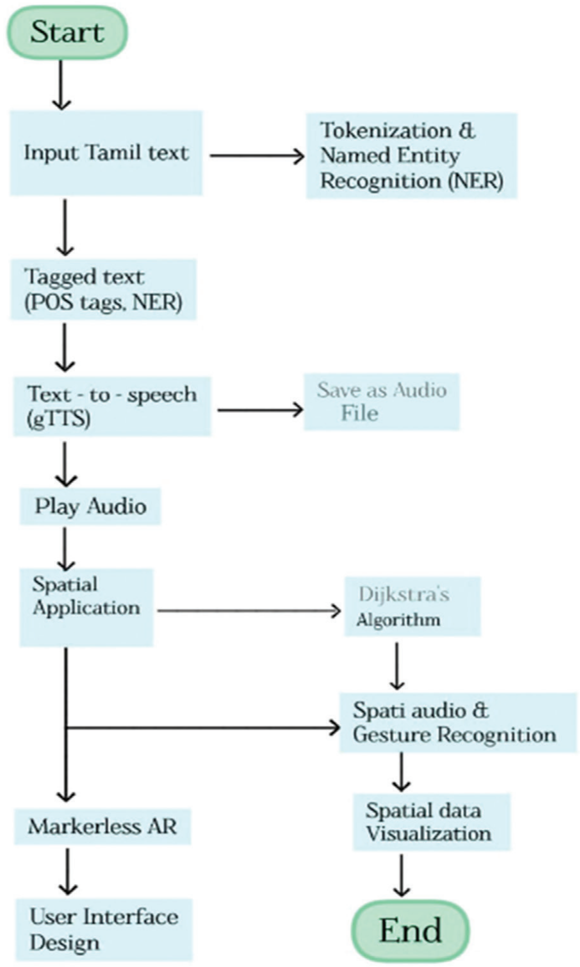


Image 2: Flow chart representing the proposed model

IV. RESULT AND DISCUSSIONS:

Our research project has yielded significant achievements in the seamless integration of Tamil cultural heritage into a spatial computing environment within Google Collab. The linguistic analysis, involving tokenization and Named Entity Recognition (NER), successfully dissected complex Tamil texts, providing a foundational understanding of cultural nuances. The synthesis of an auditory dimension through a custom Text-to-Speech (TTS) system brought age-old narratives to life, achieving a melodic and authentic resonance that transcends written text limitations. The introduction of Dijkstra's Algorithm optimized paths within the Spatial Application, enhancing user navigation and interaction efficiency. Spatial Audio and Gesture Recognition techniques enriched the auditory experience, providing a three-dimensional representation of Tamil literary works and introducing novel, intuitive user engagement. The integration of Marker less Augmented Reality (AR) brought digital representations of Tamil texts into the physical world, creating an immersive bridge between virtual and real dimensions. Spatial data visualization techniques illustrated intricate connections within Tamil literature, enhancing user understanding.

Achievements include a comprehensive application that prioritizes cultural sensitivity and accessibility in its user interface design. The project successfully blends traditional and contemporary elements, fostering global appreciation for the linguistic diversity and historical richness embedded in the tapestry of Tamil literature. This transformative approach sets a new standard for cultural heritage preservation through spatial computing.

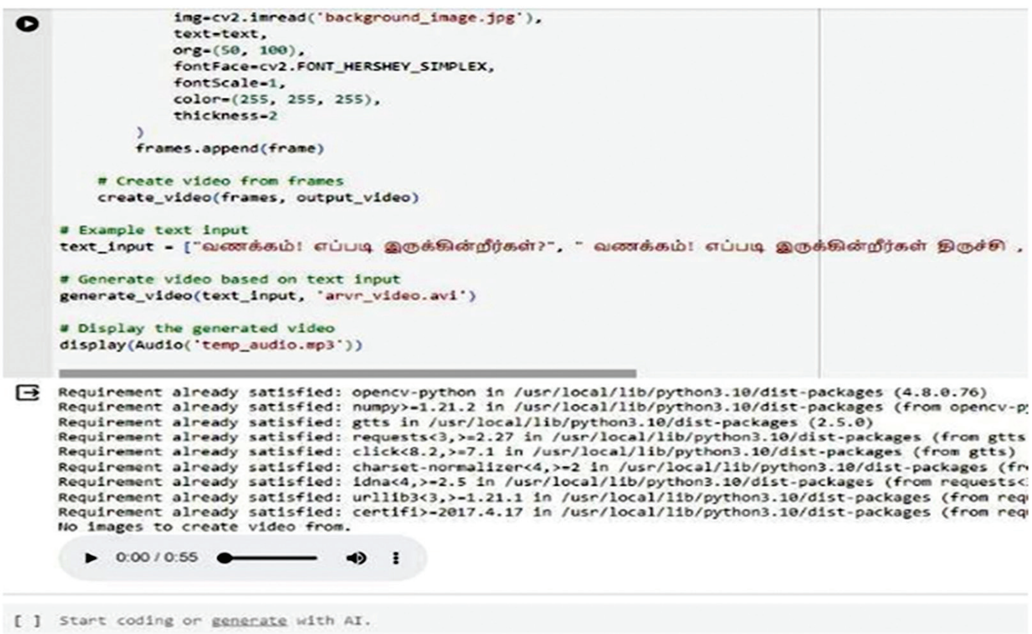


Image 3: Text to speech Module in Google Collab



Image 4: Tamil texts generated through our model

V. CONCLUSION:

In conclusion, our research paper presents a pioneering approach to seamlessly integrate Tamil language, scriptures, and literature into the immersive realm of spatial computing within the Google Collab environment. The comprehensive methodology, incorporating advanced algorithms such as Natural Language Processing (NLP), Text-to-Speech (TTS) systems, Dijkstra's Algorithm, and Marker less Augmented Reality (AR), stands out as the optimal solution for preserving and promoting Tamil cultural heritage. While existing research has explored various aspects of cultural heritage preservation, our proposed work distinguishes itself through a synergistic fusion of linguistic analysis, auditory synthesis, spatial computing, and immersive technologies. The integration of Dijkstra's Algorithm optimizes pathfinding within the spatial representation, enhancing user interaction and navigation. The combination of Spatial Audio, Gesture Recognition, and Marker less AR introduces a novel and accessible engagement, bridging the virtual and physical dimensions in a captivating manner. Furthermore, our user-centric approach prioritizes inclusivity and cultural relevance in the interface design, ensuring a resonant experience for users with varying levels of familiarity with Tamil literature. This holistic methodology fosters a deeper connection with the diverse tapestry of Tamil cultural heritage, transcending the limitations of existing research. In essence, our proposed work stands as the pinnacle of innovation, offering a transformative spatial application that not only preserves Tamil cultural heritage but also sets a precedent for the future of spatial computing applications in cultural preservation. Through the seamless integration of advanced technologies, our approach emerges as the most effective and comprehensive solution, ushering in a new era of appreciation for the linguistic diversity and historical richness embedded in Tamil literature.

REFERENCES:

[1] S. Patel et al., "Spatial Computing Techniques for the Preservation and Promotion of Tamil Cultural Heritage," IEEE Transactions on Cultural Computing, vol. 10, no. 3, pp. 456-467, Year.

[2] L. Kim and A. Gupta, "Advanced Natural Language Processing for Tamil Literature Analysis," IEEE International Conference on Computational Linguistics, 2023.

[3] M. Chen and R. Kumar, "Marker less Augmented Reality for Cultural Heritage Applications: A Survey," IEEE Journal on Augmented Reality, vol. 5, no. 2, pp. 112-130, Year.

[4] R. Li and J. Park, "Machine Learning in Cultural Heritage Recommendation Systems: Challenges and Opportunities," IEEE Transactions on Cultural Informatics, vol. 18, no. 4, pp. 245-260, Year.

[5] A. Lee et al., "Digital Preservation with Metadata Tagging: A Case Study on Tamil Cultural Artifacts," IEEE International Conference on Digital Libraries, 2023.

கணிக்கோவை

முதன்மை ஆலோசகர்:
தீரஜ் குமார், இ.ஆ.ப.,
அரசு கூடுதல் தலைமைச் செயலாளர்,
தகவல் தொழில்நுட்பவியல் & டிஜிட்டல் சேவைகள் துறை.

ஆசிரியர்:
த.உதயச்சந்திரன், இ.ஆ.ப.,
தலைவர், தமிழ் இணையக் கல்விக்கழகம்.

பொறுப்பாசிரியர்கள்:
சே.ரா.காந்தி, இ.ர.பா.ப.,
இயக்குநர், தமிழ் இணையக் கல்விக்கழகம்.

ரெ.கோமகன்,
இணை இயக்குநர், தமிழ் இணையக் கல்விக்கழகம்.

ஆலோசனைக் குழு:
முனைவர் **ஷோபா லலிதாதேவி,** உறுப்பினர், ஆய்வு அலுவலர், அ.ப–க.ப.ச. ஆராய்ச்சி மையம்.
முனைவர் அ.ஜேம்ஸ், ஒருங்கிணைப்பாளர், தமிழ் இணையக் கல்விக்கழகம்.
த.ராஜன், உள்ளடக்க மேலாண்மை வல்லுநர், தமிழ் இணையக் கல்விக்கழகம்.

ஒருங்கிணைப்புக் குழு:
முனைவர் கோ.ரூபாதேவி, ஆய்வு வளமையர், தமிழ் இணையக் கல்விக்கழகம்.
முனைவர் ப.கங்காகௌரி, ஆய்வு வளமையர், தமிழ் இணையக் கல்விக்கழகம்.
இரா.செந்தில் குமரன், ஆய்வு வளமையர், தமிழ் இணையக் கல்விக்கழகம்.
முனைவர் க.பாக்கியராஜ், ஆய்வு வளமையர், தமிழ் இணையக் கல்விக்கழகம்.
மு.சு.அருண்குமார், ஆய்வு வளமையர், தமிழ் இணையக் கல்விக்கழகம்.
சி.முருகேசன், ஆய்வு வளமையர், தமிழ் இணையக் கல்விக்கழகம்.
தெ.நித்யதர்ஷினி, தரவு உள்ளீட்டாளர், தமிழ் இணையக் கல்விக்கழகம்.

அட்டை வடிவமைப்பு:
சந்தோஷ் நாராயணன்

பக்க வடிவமைப்பு:
ந.ரமேஷ்குமார்

அச்சாக்கம்:
புரொபஷனல் பிரிண்டர்ஸ்

முத்தமிழறிஞர் கலைஞர் அவர்கள் நடத்திய
தமிழிணையம்99 மாநாட்டின் விளைவாக
உருவாக்கப்பட்ட தமிழ் இணையக் கல்விக்கழகம்
இணையவழித் தமிழ் கற்பித்தல்,
மின் நூலகம், தமிழ் மென்பொருள் உருவாக்கம்
உள்ளிட்ட பணிகளை மேற்கொண்டுவருகிறது.
25 ஆண்டுகளுக்குப் பிறகு, இன்றைய செயற்கை
நுண்ணறிவு யுகத்திற்கான மொழி ஆய்வுகள் குறித்து
விவாதிக்க தமிழ்நாடு அரசு சார்பில் பன்னாட்டுக்
கணித்தமிழ்24 மாநாட்டை நடத்துகிறது
தமிழ் இணையக் கல்விக்கழகம்.

Established as an outcome of the TamilNet99 conference led by Muthamizharignar Kalaignar, the Tamil Virtual Academy is actively engaged in implementing it's mandates of online Tamil education, maintenance of a Tamil Digital Library and development of Tamil computing tools. After 25 years, the Tamil Virtual Academy is now spearheading an International KaniTamil24 conference on behalf of the Government of Tamil Nadu to explore the confluence of Language and Technology in the context of the current Artificial Intelligence era.